DOCUMENT RESUME

ED 400 318 TM 025 714

AUTHOR Sykes, Robert C.; And Others
TITLE Scaling Polytomous Items That Have Been Scored by Two Raters.
PUB DATE Apr 96
NOTE 52p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Constructed Response; High Schools; *Item Response Theory; Mathematics Education; Multiple Choice Tests; Reading Instruction; Responses; *Scaling; Science Education; Scores; *Scoring; Standardized Tests; *Test Items
IDENTIFIERS Averaging (Mathematics); Large Scale Assessment; Partial Credit Model; *Polytomous Items; Rasch Model; *Summation (Mathematics); Three Parameter Model

ABSTRACT

The presence of multiple readings of a student response to a constructed-response item in a large-scale assessment requires a procedure for combining the ratings to obtain an item score. An alternative to the averaged item ratings that are usually used is the summing of ratings for each item. This study evaluated the effect of summing as opposed to averaging ratings in situations when both polytomous constructed-response and dichotomous selected-response (multiple choice) items were used to measure one construct and then placed on a common scale. The effects of these two aggregation methods on two item response theory models, the Rasch model, and a combination of three-parameter logistic and generalized partial credit models (the "generalized" model), were also studied. Data came from three forms of a state high school proficiency test. The effect of summing, as opposed to averaging ratings, varied across the three content areas of mathematics, reading, and science, when evaluated with the generalized model. For reading, summing reduced test information in the lower portion of the scale and increased it in the upper portions. For mathematics, the effect of summing ratings was to decrease the precision of ability estimates, and in science, summed ratings resulted in test information that was increased or decreased relative to averaged ratings in different parts of the scale. Six appendixes present supplemental information on summed ratings and item parameters for the content areas. (Contains four tables, six figures, and nine references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
ROBERT C. SYKES

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ED 400 318

Scaling Polytomous Items That Have Been Scored by Two Raters

Robert C. Sykes

Wendy Yen

Kyoko Ito

CTB/McGraw-Hill

This paper was presented at the Annual Meeting of the National Council on Measurement in Education in New York, April 1996.

TM025714

INTRODUCTION

Constructed response (c.r.) or open-ended (o.e.) items are frequently found in large scale state and national educational assessments. Student responses to these items are evaluated by one or more raters (readers) using rating scales of two or more points. Multiple readings of each constructed response are often utilized when test scores are used to make high stakes decisions about students. In these circumstances an item may be read by two readers with a third reader brought in if the initial two raters differ by more than one point.

The presence of multiple readings of a student response requires a procedure for combining the ratings to obtain an item score. Perhaps the most common means of aggregating multiple item ratings is to average the two or three readings. Frequently average scores are then rounded to the nearest integer for reporting purposes. The resulting integer item scores are "bought" at the price of as much as one half point of rounding error per item. *Individual* total raw scores from tests that contain a number of c.r. items could then contain substantially more than a half point of rounding error, although the average degree of rounding error in individual total scores would expectedly be zero.

An alternative to averaged item ratings that would reduce the amount of rounding error in individual total scores is the summing of ratings for each item. Item scores that are obtained from summed item ratings would have twice the number of total

points compared to item scores obtained from averaged, then rounded item ratings. These scores would have to be reweighted at some later time if it was desired to preserve the original relative contribution of points from o.e. items to the total score.

In cases where the initial two ratings of a student's response did not differ by more than one point - the substantial majority of the time under expected standards of inter-rater reliability for items with moderate number of levels - no rounding would be necessary. For those instances where all three readings are retained and counted toward an item score, rounding would occur after the sum of the item ratings is multiplied by 2/3. The maximum amount of rounding error in an integer item score would only be one third, as opposed to one half, of a score point, however. Thus both the frequency and magnitude of rounding error in individual total raw scores would expectedly be less for summed as opposed to averaged item ratings. It is not known to what degree an increase in precision of total scores due to decreased amounts of rounding error in summed item scores compensates for any loss in item, and possibly test, reliability due to the use of summed item ratings.

The purposes of this paper were to: 1) to evaluate the effect of summing as opposed to averaging ratings of open-ended items for the situation when both polytomous c.r. and dichotomous selected-response (multiple choice [m.c.]) items were employed to measure one construct and subsequently placed onto a common

scale, and 2) to assess the impact of these two aggregation methods on two item response theory (IRT) models. These two models were:

1) a combination of one-parameter logistic and Rasch partial credit models ("Rasch model") and
2) a combination of three-parameter logistic and generalized partial credit models ("Generalized model").

METHOD

Source Data

Source data were obtained in the form of student responses to the combination of four-choice selected response (s.r.) and c.r. items in three piloted forms of a state high school proficiency test. Items from the three forms had been successfully tried out prior to their selection for the pilot forms. One form from each of three content areas was selected: Reading, Math, and Science. The number of scored items of each type are summarized below:

	Selected Response	Open Ended
Reading	35	1
Math	40	6
Science	42	8

Each of the three samples of students contained more than 1200 students. No student took more than one form in any content area.

A confirmatory factor analysis, performed on a companion Science form using summed ratings, revealed no signs of significant multidimensionality. No corresponding analyses have

been performed for the Math or Reading forms.

Rating Process

The open-ended item in Reading had four score levels (including 0) when ratings were averaged; for Math and Science the number of levels for the open-ended items ranged between three to six and between three to four, respectively. When ratings were summed the open-ended item in Reading had seven levels, the Math open-ended items between five and 11, and the Science items between five and seven.

Each open-ended item was scored by at least two raters. Raters were trained to implement scoring rubrics, and anchor papers, check sets, and read behinds were employed to verify and maintain scoring accuracy. The smallest percentage of item readings that one of the first two raters agreed, within one point, with the second reader was 89% for Reading, 91% for Math, and 81% for Science. Consistency rates, defined as the percentage of time that any reader with at least ten reads agreed, within one point, with the second or third reader, were no lower than 90%, 93%, and 95% for Reading, Math, and Science, respectively.

Scaling Process

Multiple-choice and open-ended items were scaled together twice, using the Generalized and the Rasch models. With the Generalized model a three-parameter logistic (3pl) model (Lord,

1980) was used for the multiple-choice items.

$$P_i(\theta) = P(X_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7A_i(\theta - B_i)]} \quad (1)$$

where A_i is the discrimination, B_i is the location, and c_i is the guessing parameter for item i.

A generalization of Master's (1982) Partial Credit model (2ppc) was used for the c.r. items. For a c.r. item with m_i score levels assigned integer scores that ranged from 0 to $m_i - 1$:

$$P_{ik}(\theta) = P(X = k-1|\theta) = \frac{\exp(z_{ik})}{\sum_{l=1}^{m_i} \exp(z_{il})}, \quad k=1,\ldots,m_i \quad (2)$$

where

$$z_{il} = \alpha_i(k-1)\theta - \sum_{l=0}^{k-1}\gamma_{il},$$

and $\gamma_{i0} = 0$. α_i is the item discrimination. γ_{il} is related to the difficulty of the item levels: the trace lines for adjacent score levels intersect at γ_{il}/α_i. The parameters of the two models are related in the following manner:

$\alpha_i = 1.7A_i$ and

$\gamma_{i1} = 1.7A_iB_i$.

This Generalized model, the same as Muraki's (1992) "generalized partial credit model", will be referred to as the 3pl/2ppc model.

With the Rasch model a one-parameter model (1pl) that was obtained from (1) by setting $c_i = 0$ and $A_i = \overline{A}$ (a common discrimination value) was used for the m.c. items. Master's (1982) Partial Credit model (1ppc) was used for the open-ended items. This model is the same as the partial credit model described in (2) above with a common discrimination α substituted for the individual item discriminations, α_i. Because of the use of both the 1pl and 1ppc models, the Rasch model will also be referred to as the 1pl/1ppc model.

Parameter Estimation and Fit

Item parameter and θ estimation were conducted using the program PARDUX (Burket, 1991; 1995). Evaluations of the accuracy of the program with simulated data (Fitzpatrick, 1994) have found it to be at least as accurate as MULTILOG (Thissen, 1986) and BIGSTEPS (Linacre & Wright, 1993). Item parameters are estimated using marginal maximum likelihood procedures implemented with an EM algorithm. The ability scale is defined by specifying a prior true θ distribution to have a mean of 0.0 and standard deviation of 1.0.

Model fit was evaluated with a Fit z statistic comparing observed and predicted trace lines. The Fit z is a standardization of the Yen (1981) Q_1 statistic that facilitates comparisons of items with varying number of score levels :

$$z_{Q1} = \frac{Q_1 - df}{\sqrt{(2df)}}$$

Observed and predicted trace lines were also compared graphically. Because of the difficulty of interpreting plots for multi-level items, observed item performance was compared against predicted using the item characteristic function. The predicted function for open-ended items with m_i levels is:

$$E(X_i) = \sum_{i=1}^{mi} (k-1) P_{ik}(X_i=k-1 \mid \theta)$$

RESULTS

Raw Score Statistics

Descriptive statistics for the three forms, described for both averaged and summed item ratings, are presented in Table 1. The mean raw scores rise expectedly with the total number of points for the summed ratings, reflecting the doubling of points for the open-ended items.

Expressed as a percentage of the maximum score the three forms were moderately difficult for the students. Math was more difficult although none of the forms were so difficult as to create floor effects. The summed rating versions of the forms were more difficult than the averaged rating forms. The average percentage of maximum score correct ranged between 45% and 55% for the former and between 51% and 58% for the latter versions of the forms.

The item test correlations for all forms and versions were

similarly dispersed (standard deviations between .11 and .14) around means that varied between .40 and .45. Mean item-test corelations for the summed rating versions of forms were always slightly smaller than for the averaged rating forms.

Coefficient alphas for the summed rating forms reflected the slightly smaller average item-test correlations in being between .01 (Science) and .04 (Math) lower than the corresponding alphas for the averaged rating forms. The alphas are attenuated to some degree, probably more so for the summed ratings, because of the violation of the assumption of item tau equivalence.

Scaling Results

Fit Statistics

Table 2 contains descriptive information on Fit z statistics for the two models applied to the two types of open-ended item scores. Substantial differences in fit across the two types of models were apparent. Differences across the two types of ratings were small.

Turning first to model differences, distributions of z statistics were substantially shifted toward larger values for the 1pl/1ppc model across all three content areas with both summed and averaged item ratings. For example, in Reading 11 items had z's that were greater than or equal to ten when calibrated using the 1pl/1ppc model with summed ratings compared to none for the 3pl/2ppc model.

The largest z statistic observed for either rating type for items calibrated under the *Rasch* model was 41.60 for Math using summed ratings. The largest z statistic observed for the *Generalized* model was 12.01 for Math with averaged ratings.

The six average z statistics across the combinations of content areas and rating type ranged between 4.54 (Science:Summed) and 8.28 (Reading:Summed) for the 1pl/1ppc model versus .46 (Math:Averaged) to 1.25 (Reading:Summed) for the 3pl/2ppc model.

The value of z tends to increase with sample size so the statistic is typically compared to critical values that increase with the size of samples. A flagging criterion of approximately 4.0 has been used by CTB for samples of the size available for this study.

The number of z statistics greater than 4.0 across the three content areas and two types of aggregated ratings (found in the last column of Table 2) ranged between 18 (Math:Averaged) and 25 (Math:Summed) for the 1pl/1ppc model. The range of z's greater than four for the 3pl/2ppc model ranged between zero (Reading:Summed) and two (Science:Summed).

Turning to fit differences across the two types of aggregated ratings, the fit of the *Rasch* model was better, in terms of the average z statistic and fewer z's above 4.0, with averaged as opposed to summed ratings in Reading and Math. The greatest improvement in fit occurred for Math (25 vs. 18 misfitting items for summed vs. averaged ratings, respectively).

The substantial 39% of the Math items that failed to fit the 1pl/1ppc model with averaged ratings (18/46) was the smallest proportion of misfitting items for the three content areas. The number of open-ended items that were not fitted by the Rasch model with averaged versus summed ratings was 0,1; 3,5; and 3,2 for Reading, Math, and Science, respectively.

Plots of fit for selected items calibrated with the two models and two types of aggregated ratings are presented in Appendices. Appendices A1, B1, and C1 depict one selected m.c. and open-ended item for Reading, Math, and Science (respectively) that were calibrated with the 1pl/1ppc model using averaged ratings. To the right of these plots are the same items fitted with the 3pl/2ppc model. Appendices A2, B2, and C2 contain two plots for the same selected open-ended item in the three content areas when calibrated with the 1pl/1ppc and 3pl/2ppc models with summed ratings.

The three selected m.c. items from the three content areas demonstrate the 1pl/1ppc model's difficulty dealing with the effect of guessing (Reading:Item 20 and Science:Item 23) as well as differences in discrimination (Math:Item 34 and Science:Item 23). The open-ended items reveal a tendency for the Rasch model to better fit averaged, as opposed to summed, ratings.

Item Discriminations

Item parameter estimates for Reading are presented for both types of ratings and the two models in Appendix D. Parameter

estimates for Math and Science may be found in Appendix E and F, respectively.

With the 3pl/2ppc model, item discrimination did not practically change between summing and averaging for the m.c. items. This is consistent with the fact that the m.c. items did not involve summing or averaging. For the o.e. items the 3pl/2ppc discriminations were considerably greater for averaged (fewer score levels) than for summed ratings (more score levels). This indicates that the greater the number of score levels, the lower the discrimination because it becomes increasingly difficult to differentiate between individual levels.

With the 1pl/1ppc model, which fixes a constant average discrimination for all items, the discrimination was always higher for averaging than for summing. Thus, the 1pl/1ppc model presents a picture that contradicts the fact that m.c. items did not involve averaging or summing.

Discrimination estimates obtained using the 3pl/2ppc model are markedly lower for the o.e. items than for the m.c. items. For *averaged* ratings in Reading, Math, and Science, average discriminations were 1.50, 1.63, and 1.32 for the m.c. items, respectively, versus .74, .88, and 1.10 for the o.e. items. The dispersion of discrimination estimates was also larger for the m.c. items in Math and Science (.47 and .56) relative to the dispersion of the o.e. items in these content areas (.24 and .28). The difference between average m.c. discriminations for the *summed* ratings (1.51, 1.64, and 1.31 for Reading, Math, and

Science, respectively) and average o.e. discriminations for the three forms (.39, .44, and .61) is even larger.

The approximate 50% drop in the average discrimination estimates for *open-ended* items calibrated with summed ratings relative to the averages for averaged ratings reflect the approximate doubling of the number of levels when ratings are summed rather than averaged. The progressive decrease in discrimination parameter estimates from m.c. items to o.e. items with averaged ratings to o.e. items with summed ratings is, thus, to be expected. The capability of an item to place an examinee in one category/level versus another should decrease as the total number of categories within which the examinee could potentially be placed increases.

The systematic differences in discrimination were thus consistent with the model fit results. The Rasch model, which assumes all items have the same capability to place examinees in one category versus another, did not accurately describe performance on items that differed in their number of levels nor items that differed in their capability to place an examinee in one of only two categories. While the model could be applied to a subset of m.c. items in the latter category that were screened for a restricted range of discrimination, screening could not be expected to eliminate the problem of misfit with mixed item type tests containing o.e. items with larger number of levels. In this case open-ended items would have to be eliminated solely because of their greater number of score levels.

Item Information

The lower discriminations for the open-ended items using either averaged or summed ratings does not result in these items contributing less information to the overall test score. Item information is related to item discrimination in the following manner:

$$I(X_i) = \alpha_i^2 \rho^2(X_i|\theta)$$

Because item information is a function of both item discrimination and the variance of the item scores, items with greater score variance can have larger amounts of information, even with smaller discriminations. Consequently o.e. items, with their greater number of score levels, generally provide more information than m.c. items.

The Rasch model requires that all items be described as having the same discrimination; hence items with more score levels must necessarily be described as providing more information. Furthermore, since item scores for summed ratings have more levels than item scores for averaged ratings, the Rasch model must describe open-ended items scored by summing ratings as providing very high levels of information.

The total information provided by each item (multiplied by 1000) under the Rasch and the Generalized model is provided for the items in the three content area forms when item ratings are averaged in Table 3. (Total information is defined as the area under the item information function obtained after item

parameters have been transformed to a scale score metric with mean of 500 and standard deviation of 50.) Table 4 contains the corresponding figures when item ratings are summed.

In Table 3 Math o.e. items #21 and #25 have Rasch information values of 801 and 1189, respectively. Both of these maximum four point items demonstrated substantial misfit to the 1pl/1ppc model yet were fitted by the 3pl/2ppc model. (The only Math item demonstrating substantial misfit under the 3pl/2ppc model with either type of rating was the m.c. item #4.) The corresponding total information for these items under the 3pl/2ppc model is less than 1/2 the total information predicted by the 1pl/1ppc model. Although the item that contributes the most total information (under the 3pl/2ppc model) is a multiple choice item (#34 with 632), the average total item information contributed by the Math o.e. items (492) greatly exceeds the average information contributed by the m.c. items (196).

The results for the summed ratings in Table 4 indicate that Math items #21 and #25, misfitting again under the 1pl/1ppc model and fitted again by the 3pl/2ppc model, have Rasch total information four to six times that of the items' total information under the 3pl/2ppc model. Rasch information values for the o.e. items have increased substantially with the summed ratings. The average Rasch information across all 15 o.e. items in the three content areas was 955 compared to 420 for the same items using averaged ratings. The average 3pl/2ppc item information contributed by the 15 o.e. items increased to a much

smaller degree between averaged and summed ratings. The average information under the 3pl/2ppc model for averaged ratings was 372 versus 425 for summed ratings.

Test Characteristics

The large percentage of items that misfit when either averaged or summed ratings were used with the Rasch model (a minimum of 39%), in conjunction with the spurious substantially large total item information values produced for the o.e. items, suggests the possibility that model misspecification error could impact the measurement characteristics of form scores.

To evaluate this possibility test characteristics curves (tcc's) were obtained for both models using both types of aggregated ratings. The tcc's are plotted with score information functions obtained for the 1pl/1ppc model and (unweighted) number-correct scores using the 3pl/2ppc item parameter estimates. Figure 1a, 2a, and 3a contain these plots for averaged o.e. ratings from the Reading, Math, and Science forms, respectively. Figures 1b, 2b, and 3b contain the corresponding plots for the summed ratings. A distribution of maximum likelihood ability estimates from number correct scores, obtained from a generalization of a procedure for m.c. tests by Yen (1984), is plotted at the top of each figure. Finally each figure contains a tabular display of the relationship between scale scores and raw scores at consecutive scale score intervals and standard errors of number correct ability estimates, obtained

as the squared root of the reciprocal of score information.

Substantial differences in the lower portions of the 1pl/1ppc versus 3pl/2ppc tcc's for *averaged* ratings are present in each content area. Predicted raw scores for the 3pl/2ppc model exceed those for the 1pl/1ppc model by more than one and one half raw score points up through a scale score of 400 for Reading and Math. For Math the difference is as large as 2.41 raw score points at 400 (10.80 for 1pl/1ppc vs 13.21 for 3pl/2ppc). The differences between the lower portions of the tcc's diminish with the *summed* ratings for Math and Science plotted in Figures 2b and 3b. Differences between predicted raw scores are less than one raw score point at a scale score of 400 for all content areas.

The substantially better fit demonstrated by the 3pl/2ppc model and the failure of the 1pl/1ppc to adequately account for guessing in these somewhat difficult forms suggest that the 1pl/1ppc tcc predicts an erroneously low raw score throughout the lower 25% of test scores that is still relatively large at a scale score of 400 for averaged ratings. The seriousness of this underprediction would depend to a large degree on whether a cutscore expressed on the scale score scale fell within this region.

Agreement between 1pl/1ppc and 3pl/2ppc tcc's is better above 400 for all three content areas with both types of item scores. With the exception of a 2.24 raw score difference at a scale score of 575 for Math summed scores (69.08:1pl/1ppc vs

66.84:3pl/2ppc), all differences are less than 1.5 raw score points.

The plots of 1pl/1ppc and 3pl/2ppc score information functions are provided in the six figures along with the information obtained from item pattern scoring. The pattern scoring information function serves as a reference point, indicating the maximum possible information that could be obtained when the item response data is generated by a 3pl/2ppc model.

An evaluation of the score information plots indicates that the 1pl/1ppc model spuriously inflates its prediction of information present throughout the lower range of the scale up to scale scores of approximately 425 or 450. This effect, due to the model's failure to account for guessing, results in the 1pl/1ppc standard error of ability estimates being substantially attenuated (relative to the number correct 3pl/2ppc standard errors) in this lower range of the scale for both types of aggregated ratings across all three content areas.

Across all six combinations of content areas by rating type the difference in the model standard errors (3pl/2ppc - 1pl/1ppc) at a scale score of 400 ranges between 11.89 for Science:Summed scores to 29.50 for Reading:Summed scores. This difference is a sizable 24% to 59% of a scale score standard deviation. The frequency distributions of number correct 3pl/2ppc scale scores for each of the six combinations indicate that non-negligible numbers of students fall within this lower portion of the scale

and consequently would have sizably underestimated standard errors if the 1pl/1ppc was used for scoring.

The effect of the spurious inflation of the amount of *item* information predicted by the Rasch model for o.e. items with large numbers of levels, such as what occurs for the summed ratings for Math mentioned earlier, is noticeable in Figure 2b. The amount of score information predicted by the Rasch model at scale scores between 525 and 575 actually *exceeds* the information predicted under pattern scoring. The inflated score information results in a 1pl/1ppc ability standard error at a scale score of 550 that is more than four points less than the 3pl/2ppc standard error (10.82 vs 14.96). Inspection of the frequency distribution indicates that the apparent underestimation of the error in ability estimates at this interval of the scale would impact substantial numbers of examinees.

Average vs Summed Ratings: Generalized Model

Figures 4, 5, and 6 contain score information functions and scale score-to-raw score tables for the 3pl/2ppc model fitted on both types of ratings for Reading, Math, and Science, respectively. The effect of summing, as opposed to averaging, ratings differs for the three content areas.

With the relatively short 36 item Reading form the relative efficiency (ratio of information functions) of summed scores, with respect to averaged scores, is less than one up to a scale score of approximately 450 - 475. The relative efficiency of

summed scores exceeds one from this point through 575, again slipping below one above this point. Changing from averaged to summed scores for the one Reading o.e. item has a noticeable effect on the information of the test, and hence the error characteristics, through approximately the lower half of the score range. The difference in the standard error of ability estimates has narrowed to less than one scale score point by a scale score of 475.

The effect of summing versus averaging ratings for the larger 46 item Math form, with its greater number (6) of o.e. items having large numbers of levels, is one of reducing test information across virtually the entire score range. The large number of levels of the o.e. items, as large as 11 when summed ratings are utilized, reduces the precision (increases the standard error) of total scores or scores derived from them, relative to average ratings, at and above a scale score of 350. The additional levels created from summed ratings, reflecting degree of rater agreement (e.g. the only way to obtain an item score of one is to obtain a rating of one from one rater and 0 from the other), do not, with the many-leveled Math items, contribute to the differentiation of more able students on the primary ability.

Finally, the effect of summing ratings for the eight o.e. items in the larger, 50 item Science form with their fewer yet more similar number of levels, is mixed. Summing ratings increases information in the lower and upper part of the score

range - from the floor through 450 and from 575 to the ceiling of the test. The information produced with summed scores is slightly less than that with averaged scores in the middle of the score range. The increase in test information at the lower end of the scale for summed ratings is counter to the reduced test information noticed for Reading in this portion of the scale and Math from 350 to 450. It suggests that for the measurement of lower ability students in Science, as opposed to Math and Reading, more precision is gained by adding score levels than is lost through greater inaccuracy in item scores. More work needs to be done to confirm that the effect is not due to the relative location of the o.e. items, as opposed to their number of levels.

CONCLUSIONS

1) Substantially more items did not fit the Rasch model as compared to the Generalized model when item response data from mixed item type forms was analyzed. No fewer than 39% of the items failed to fit a combination 1pl/1ppc model across three content area forms and two types of aggregated ratings compared to no more than 4% of the items not fitting the 3pl/2ppc model. The greater degree of misfit with the Rasch model may be ascribed to the model's failure to allow for guessing and differences in item discrimination, the latter enhanced by the multi-level o.e. items.

2) Of the two content areas that had substantial numbers of o.e. items, the number of items misfitting under the 1pl/1ppc model decreased from summed to averaged ratings for Math and remained approximately the same for Science. Math was the content area having o.e. items with the most number of levels. The o.e. items in Science, while the most numerous (8) of the three content area forms, had relatively few levels and the number of these levels differed by no more than one. This suggest that the difficulty for the Rasch model was fitting a set of items that *varied* greatly in terms of number of levels, rather than the o.e. items having a large number of levels. Tests consisting of all o.e. items with a similar large number of levels might be expected to be better fit by the Rasch model.

3) The magnitude of discrimination parameter estimates decreased as the number of levels of the o.e. items increased when scaled using the Generalized model. The effect of item discriminations that decrease with increasing numbers of score levels results in the spurious inflation of item information for models, such as the 1pl/1ppc, that fix item discrimination at a constant value.

4) Differences between 1pl/1ppc and 3pl/2ppc test characteristic curves exceeded 1 1/2 raw score points through the lower portion of the score range for all three content areas. With averaged ratings the 1pl/1ppc model predicted a raw score at a scale score of 400 that was as much as 2.41 raw score points below the raw

score predicted by the 3pl/2ppc model (Math). Differences between tcc's were generally smaller for summed ratings for all three content areas.

5) Standard error of ability estimates were substantially underpredicted (24% to 59% of a scale score standard deviation) for the 1pl/1ppc model relative to number correct 3pl/2ppc standard errors throughout the lower 1/4 of the score range.

6) Spuriously inflated predictions of the amount of total item information under the 1pl/1ppc model results in ability standard errors that are attenuated by as much as approximately four points at the middle of the score range.

7) The effect of summing, as opposed to averaging ratings, varied across the three content areas when evaluated with the Generalized model. With the smallest content area form, Reading, summing ratings for the single o.e. item substantially reduced test information through the lower portion of the score range although it produced increased information in upper portions of the scale. The effect of summing ratings for the many-leveled Math items was to decrease the precision of ability estimates throughout most of the scale score range. Finally, in Science summed ratings resulted in test information that was, in various parts of the score range, increased or decreased relative to that for averaged ratings.

References

Burket, G. R. (1991; 1995). PARDUX. Monterey, CA: CTB/McGraw-Hill.

Fitzpatrick, A. R. (1994). Two studies compaing parameter estimates produced by PARDUX and BIGSTEPS. Unpublished manuscript.

Linacre, J. M. & Wright, B. D. (1993). A user's guide to Bigsteps: Rasch model computer program. Chicago: MESA.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum associates.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

Thissen, D. (1986). MULTILOG: Multiple categorical item analysis and test scoring, Version 5. Mooresville, IN: Scientific Software.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. Journal of Educational Measurement, 21, 93-111.

Table 1

Descriptive Statistics

Rating	# Scored Items	# Points	Mean	% of Max Score[1]	s.d.	N	∝	p-value[2] Mean	p-value[2] s.d.	Item - Test Correlation Mean	Item - Test Correlation s.d.
					Reading						
Averaged	36	38	21.96	58	7.23	1448	.86	.58	.19	.42	.11
Summed	36	41	22.55	55	7.73	1469	.84	.59	.16	.41	.12
					Math						
Averaged	46	61	31.18	51	12.82	1217	.91	.57	.19	.45	.13
Summed	46	82	36.93	45	18.07	1217	.87	.57	.19	.43	.13
					Science						
Averaged	50	60	32.99	55	11.45	1293	.90	.57	.16	.41	.12
Summed	50	78	39.07	50	15.28	1293	.89	.57	.16	.40	.14

[1] Mean divided by maximum score (percentage of maximum score).
[2] P-value for open-ended items computed as average score divided by the maximum score.

Table 2

Fit Statistics

Averaged Ratings

	Form	N	# of Scored Items	Fit Zs' -2≤Z<2	2≤Z<4	4≤Z<6	6≤Z<10	10≤Z	Two Largest Z's		Average Z	# Misfitting Items Z>4.0
Reading												
1PL/1PPC	2	1448	36	9	6	3	8	10	35.65	28.69	7.87	21
3PL/2PPC	2	1448	36	26	9	1	0	0	4.07	3.43	1.10	1
Math												
1PL/1PPC	6	1217	46	19	9	6	5	7	36.64	17.88	4.92	18
3PL/2PPC	6	1217	46	41	4	0	0	1	12.01	3.56	0.46	1
Science												
1PL/1PPC	14	1293	50	20	9	8	6	7	25.51	20.45	5.03	21
3PL/2PPC	14	1293	50	44	5	1	0	0	4.73	2.92	0.68	1

Summed Ratings

	Form	N	# of Scored Items	Fit Zs' -2≤Z<2	2≤Z<4	4≤Z<6	6≤Z<10	10≤Z	Two Largest Z's		Average Z	# Misfitting Items Z>4.0
Reading												
1PL/1PPC	2	1448	36	8	6	3	8	11	32.85	27.02	8.28	22
3PL/2PPC	2	1448	36	26	10	0	0	0	3.76	3.47	1.25	0
Math												
1PL/1PPC	6	1217	46	13	8	4	12	9	41.60	31.38	7.20	25
3PL/2PPC	6	1217	46	42	3	0	0	1	10.44	3.36	0.55	1
Science												
1PL/1PPC	14	1293	50	21	8	6	7	8	16.34	14.40	4.54	21
3PL/2PPC	14	1293	50	45	3	1	1	0	6.52	5.22	0.86	2

Table 3
Item Discrimination and Item Information
Averaged Ratings

	Reading			Math			Science		
	3PL/2PPC		Rash	3PL/2PPC		Rash	3PL/2PPC		Rash
Item	Discrim.	Info.	Info.	Discrim.	Info.	Info.	Discrim.	Info.	Info.
1	1.27	93	81	1.07	75	93	0.89	49	79
2	0.98	57	81	1.29	110	93	0.98	58	79
3	1.70	163	81	1.87	235	93	1.15	84	79
4	2.10	275	81	1.57	165	93	1.15	78	79
5	1.10	76	81	0.89	54	93	1.34	110	79
6	1.15	82	81	1.10	80	93	1.28	91	79
7	1.15	73	81	1.27	109	93	0.86	45	79
8	0.87	46	81	1.31	108	93	2.14	312	79
9	1.20	88	81	2.14	251	93	0.86	42	79
10	1.00	63	81	2.72	573	93	1.73	152	79
11	2.20	245	81	1.28	110	93	1.18	75	79
12	1.41	115	81	2.47	373	93	1.05	62	79
13	0.66	27	81	1.98	237	93	1.23	88	79
14	0.77	44	81	1.42	126	93	0.69 (3)	113	205
15	1.41	113	81	1.81	223	93	1.26 (3)	331	202
16	2.76	495	81	1.26	105	93	1.27 (3)	344	205
17	0.98	58	81	1.70	194	93	1.38	106	79
18	0.80	40	81	1.20	96	93	1.25	89	79
19	2.20	284	81	1.33	119	93	0.88	46	79
20	2.18	292	81	1.08	75	93	0.96	60	79
21	2.56	352	81	0.64 (5)	378	801	1.61	148	79
22	1.76	175	81	0.85 (6)	555	688	0.82	41	79
23	1.18	69	81	0.91 (4)	563	628	1.21	77	79
24	1.19	75	81	1.08 (4)	584	489	0.67	25	79
25	1.83	195	81	0.59 (5)	460	1189	2.02	252	79
26	3.00	540	81	1.20 (3)	412	286	2.28	341	79
27	1.19	77	81	0.87	50	93	0.86	163	79
28	1.57	146	81	1.37	124	93	1.22	76	79
29	0.69	30	81	0.91	56	93	1.52	143	79
30	2.06	267	81	1.36	125	93	1.07 (4)	382	317
31	0.37	9	81	0.38	10	93	1.35 (3)	307	168
32	1.71	169	81	1.56	167	93	0.76 (3)	187	282
33	1.53	134	81	1.79	209	93	1.60	145	79
34	1.66	164	81	2.99	632	93	0.85	37	79
35	2.20	300	81	1.96	226	93	1.34	134	79
36	0.74 (4)	235	360	2.53	438	93	1.14	79	79
37				2.65	502	93	2.03	216	79
38				1.65	170	93	0.77	35	79
39				1.52	146	93	1.92	193	79
40				1.91	247	93	1.26	61	79
41				1.82	203	93	2.19	281	79
42				1.77	205	93	0.51	15	79
43				1.63	170	93	1.54	133	79
44				2.50	393	93	1.12	74	79
45				1.55	162	93	2.28	276	79
46				1.62	171	93	1.43 (4)	566	305
47							0.90	48	79
48							1.52	129	79
49							1.88	186	79
50							0.93 (3)	163	173

(#): number of levels (including 0) of open-ended item
Info. and Rasch Info multiplied by 1000

BEST COPY AVAILABLE

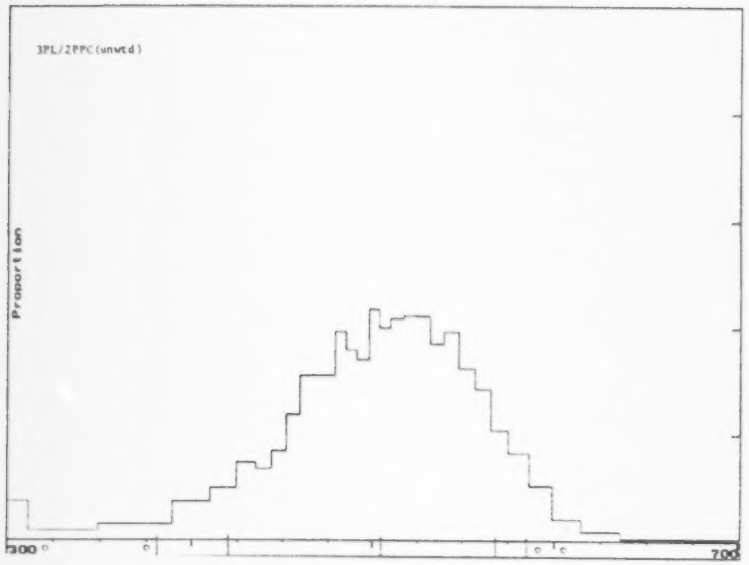
Table 4
Item Discrimination and Item Information
Summed Ratings

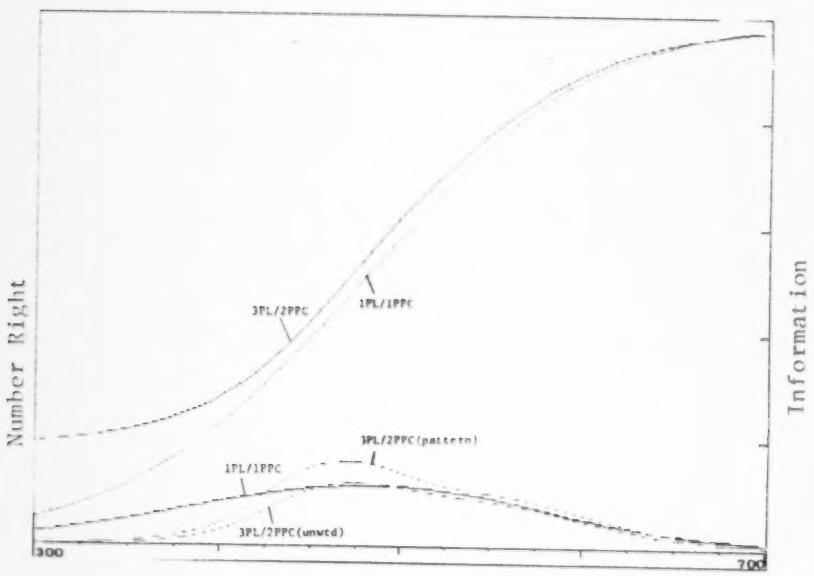
	Reading			Math			Science		
	3PL/2PPC		Rash	3PL/2PPC		Rash	3PL/2PPC		Rash
Item	Discrim.	Info.	Info.	Discrim.	Info.	Info.	Discrim.	Info.	Info.
1	1.27	100	73	1.07	78	54	0.88	52	53
2	0.98	66	73	1.28	112	54	0.99	67	53
3	1.74	190	73	1.88	241	54	1.15	91	53
4	2.28	325	73	1.56	166	54	1.13	88	53
5	1.11	85	73	0.88	53	54	1.35	124	53
6	1.15	90	73	1.10	82	54	1.26	82	53
7	1.16	88	73	1.25	107	54	0.84	48	53
8	0.88	53	73	1.32	104	54	2.14	281	53
9	1.20	98	73	2.22	252	54	0.87	49	53
10	1.01	69	73	2.76	657	54	1.72	195	53
11	2.35	286	73	1.28	112	54	1.18	94	53
12	1.41	136	73	2.52	388	54	1.05	76	53
13	0.66	30	73	1.98	234	54	1.23	102	53
14	0.74	34	73	1.45	137	54	0.42 (5)	179	479
15	1.42	137	73	1.80	221	54	0.69 (5)	403	433
16	2.75	515	73	1.30	107	54	0.70 (5)	419	447
17	0.99	66	73	1.69	196	54	1.37	128	53
18	0.81	45	73	1.20	98	54	1.24	104	53
19	2.21	327	73	1.43	124	54	0.88	52	53
20	2.30	374	73	1.08	79	54	0.94	60	53
21	2.59	373	73	0.32 (9)	392	1697	1.57	163	53
22	1.76	211	73	0.41 (11)	612	1623	0.81	45	53
23	1.13	89	73	0.41 (7)	514	1540	1.11	74	53
24	1.20	98	73	0.53 (7)	585	1001	0.66	29	53
25	1.83	228	73	0.28 (9)	453	2735	2.14	288	53
26	3.00	608	73	0.70 (5)	538	585	2.18	298	53
27	1.20	97	73	0.86	50	54	0.84	48	53
28	1.57	168	73	1.38	130	54	1.19	98	53
29	0.69	33	73	0.92	58	54	1.51	156	53
30	2.05	305	73	1.33	123	54	0.56 (7)	471	723
31	0.38	10	73	0.38	10	54	0.74 (5)	352	341
32	1.72	202	73	1.57	173	54	0.41 (5)	225	684
33	1.54	164	73	1.83	220	54	1.57	167	53
34	1.68	177	73	3.00	662	54	0.84	48	53
35	2.21	341	73	2.00	227	54	1.38	109	53
36	0.39 (7)	288	1081	2.51	443	54	1.15	90	53
37				2.65	525	54	2.00	247	53
38				1.69	156	54	0.77	40	53
39				1.55	152	54	1.85	233	53
40				1.93	262	54	1.25	107	53
41				1.81	189	54	2.16	282	53
42				1.79	218	54	0.51	18	53
43				1.65	180	54	1.52	158	53
44				2.67	465	54	1.12	85	53
45				1.55	164	54	2.24	324	53
46				1.65	176	54	0.82 (7)	762	629
47							0.90	55	53
48							1.50	146	53
49							1.88	207	53
50							0.51 (5)	180	324

(#): number of levels (including 0) of open-ended item
Info. and Rasch Info. multiplied by 1000.

BEST COPY AVAILABLE

Figure 1a
Reading: Averaged Scores

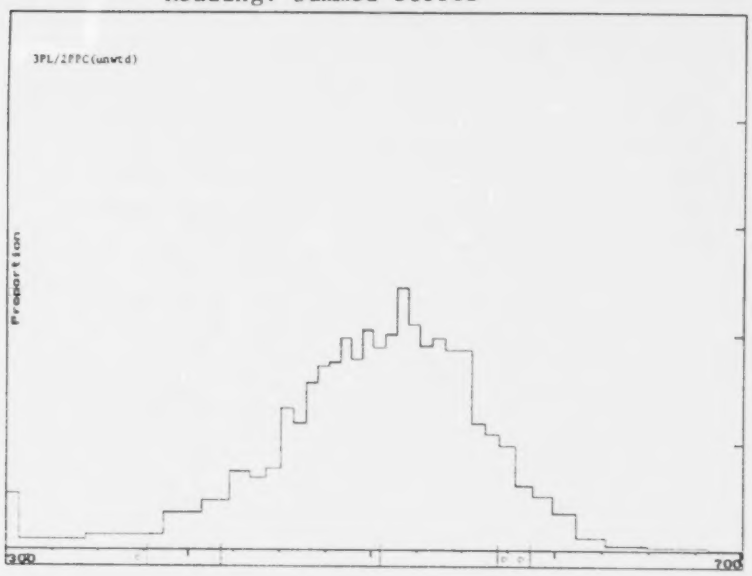


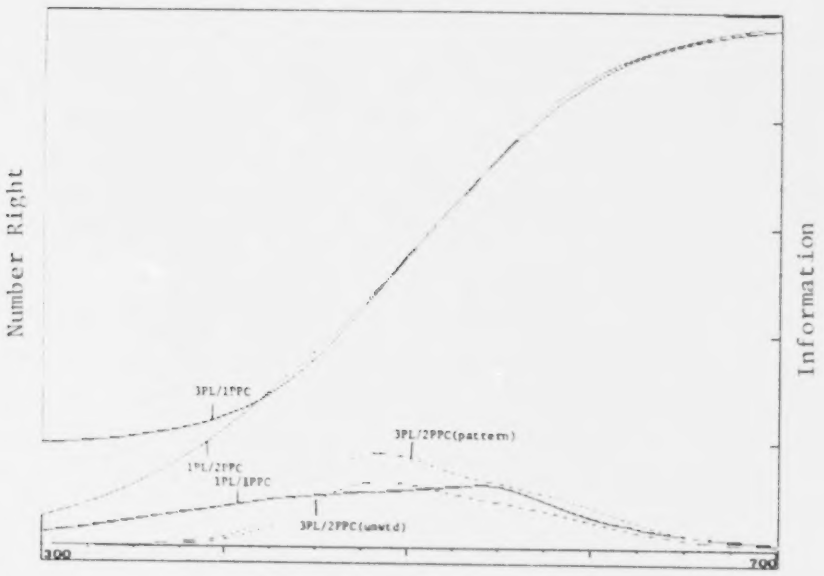


Scale Score	1pl/1ppc Raw Score	s.e.	3pl/2ppc(unwtd) Raw Score	s.e.
300	2.00	41.10	7.31	211.50
325	2.99	34.32	7.66	136.94
350	4.39	29.18	8.22	88.35
375	6.28	25.40	9.12	56.70
400	8.70	22.73	10.58	36.40
425	11.63	20.96	12.92	24.86
450	14.98	19.94	16.18	20.10
475	18.56	19.55	19.89	19.03
500	22.18	19.69	23.52	19.81
525	25.67	20.37	26.76	21.16
550	28.82	21.78	29.61	22.48
575	31.47	24.31	32.02	24.75
600	33.52	28.22	33.89	28.99
625	34.99	33.65	35.21	36.06
650	36.01	40.76	36.09	46.49
675	36.70	49.90	36.65	59.99
700	37.16	61.51	37.02	76.34

BEST COPY AVAILABLE

Figure 1b
Reading: Summed Scores

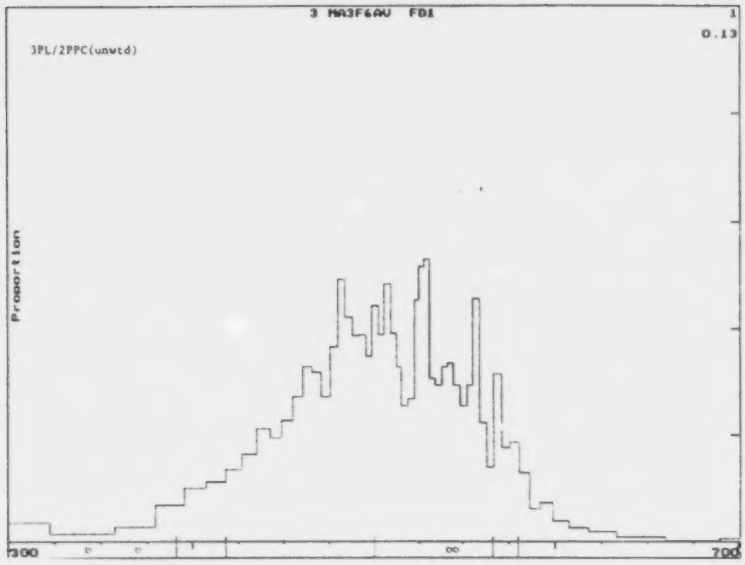


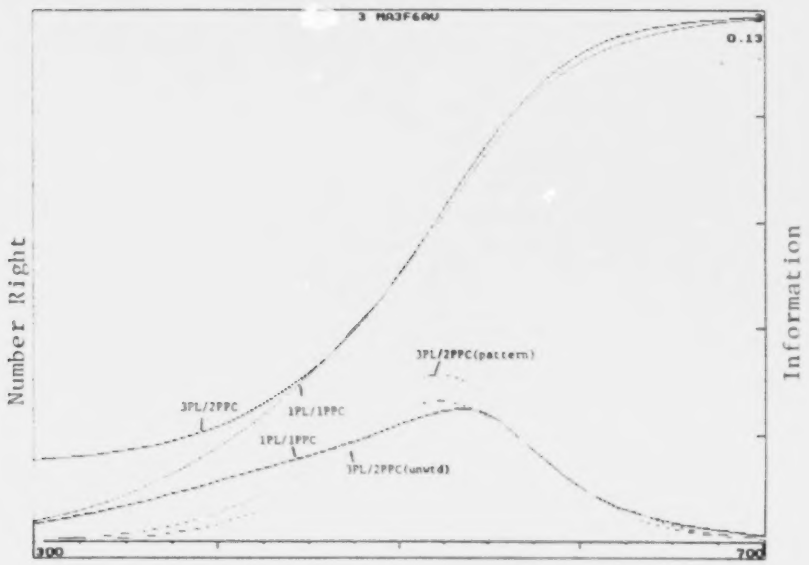


Scale Score	1pl/1ppc Raw Score	s.e.	3pl/2ppc(unwtd) Raw Score	s.e.
300	2.31	40.38	7.90	335.58
325	3.37	34.18	8.13	216.44
350	4.83	29.47	8.50	137.94
375	6.75	25.98	9.09	86.54
400	9.16	23.49	10.08	52.99
426	12.01	21.84	11.77	32.36
450	15.23	20.85	14.55	22.13
475	18.69	20.31	18.32	18.78
500	22.29	19.88	22.45	18.84
525	26.11	19.17	26.40	20.42
550	30.14	19.25	30.01	22.24
575	33.65	22.10	33.21	24.74
600	36.10	27.16	35.74	29.16
625	37.72	33.30	37.51	36.78
650	38.81	40.55	38.63	48.08
675	39.54	49.39	39.33	62.24
700	40.04	60.35	39.78	78.62

Figure 2a

Math: Averaged Scores

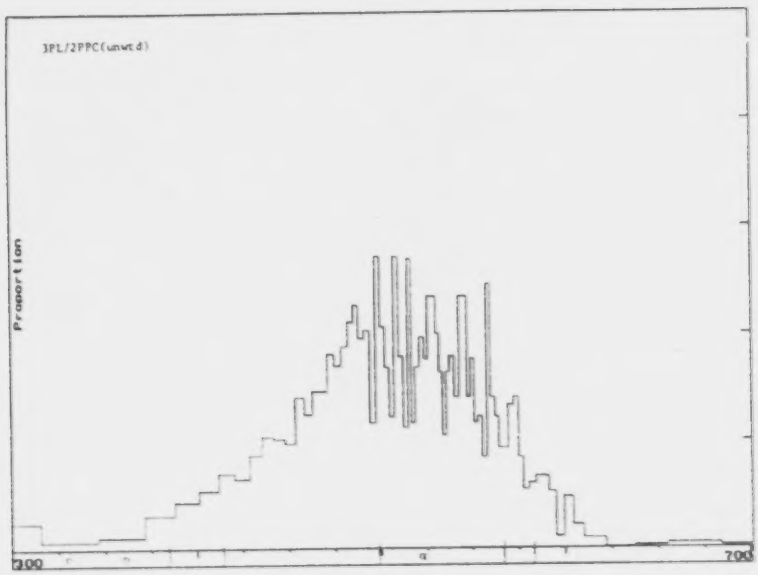


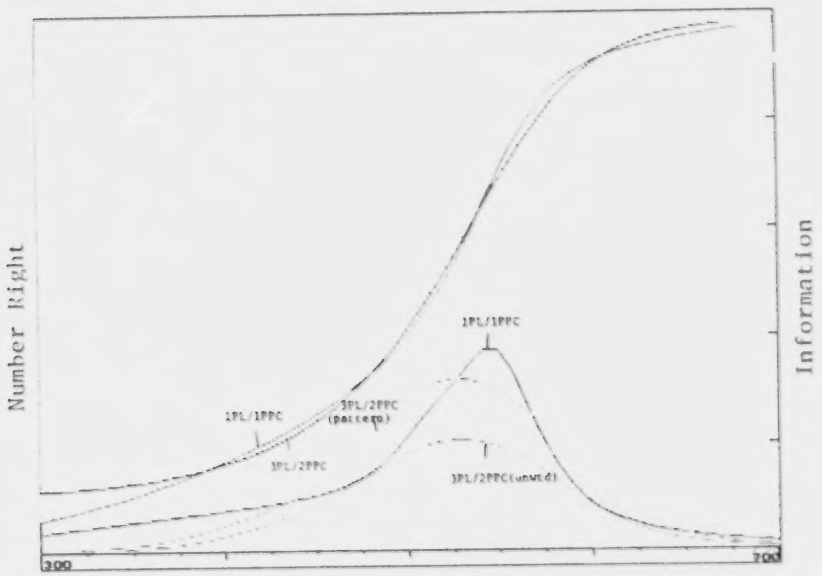


Scale Score	1pl/1ppc Raw Score	s.e.	3pl/2ppc(unwtd) Raw Score	s.e.
300	2.38	35.51	9.30	208.32
325	3.62	29.57	9.72	129.28
350	5.38	25.14	10.39	80.74
375	7.74	21.87	11.49	51.99
400	10.80	19.47	13.21	35.58
425	14.57	17.70	15.73	26.07
450	19.05	16.34	19.26	19.86
475	24.29	15.11	24.13	15.66
500	30.44	13.96	30.56	13.36
525	37.48	13.24	38.13	12.74
550	44.90	13.45	45.66	13.59
575	51.08	16.03	51.81	15.87
600	55.03	20.73	55.92	20.41
625	57.37	27.03	58.21	28.74
650	58.75	34.79	59.37	41.58
675	59.60	44.23	59.97	59.25
700	60.13	55.95	60.30	81.94

BEST COPY AVAILABLE

Figure 2b
Math: Summed Scores

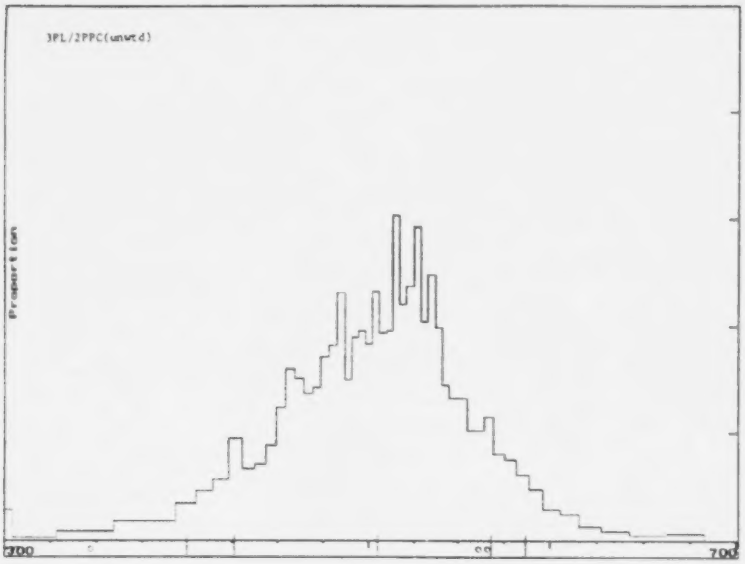


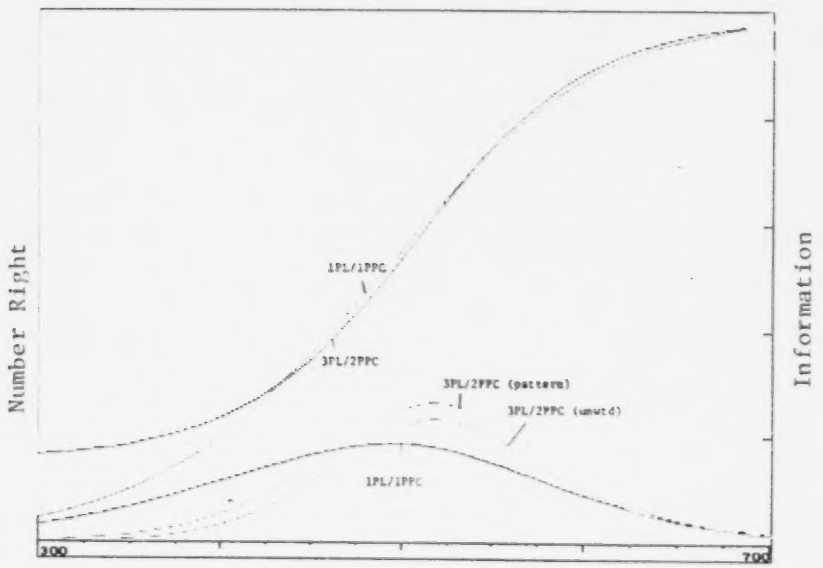


Scale Score	1pl/1ppc Raw Score	s.e.	3pl/2ppc(unwtd) Raw Score	s.e.
300	4.99	34.16	9.46	199.87
325	6.64	30.57	9.89	128.11
350	8.66	27.72	10.59	82.05
375	11.09	25.46	11.71	53.80
400	13.93	23.60	13.48	37.32
425	17.23	21.93	16.14	27.80
450	21.09	20.11	20.01	21.75
475	25.89	17.51	25.68	17.67
500	32.86	13.98	33.85	15.22
525	43.57	11.63	44.59	14.33
550	57.84	10.82	58.47	14.96
575	69.08	14.95	66.84	17.35
600	74.28	22.34	73.83	22.57
625	76.83	29.84	77.56	32.10
650	78.38	38.71	79.37	46.13
675	79.45	43.91	80.28	64.06
700	80.19	52.13	80.79	85.15

Figure 3a

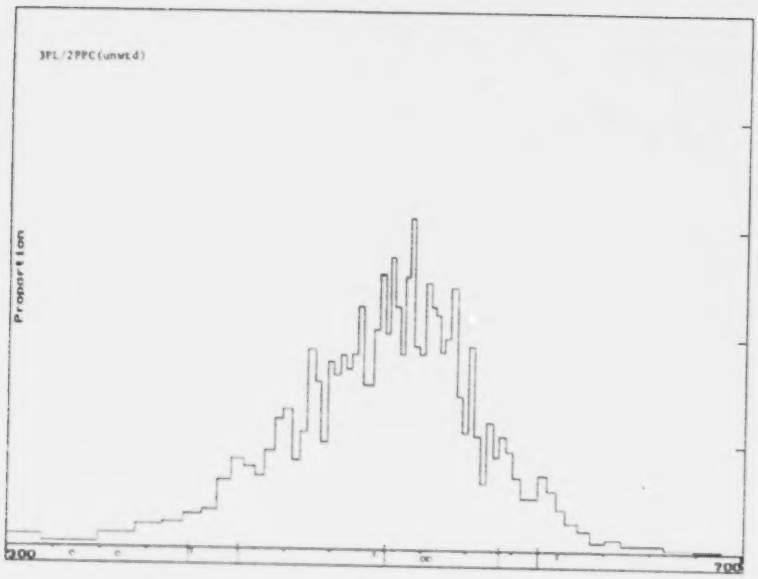
Science: Averaged Scores

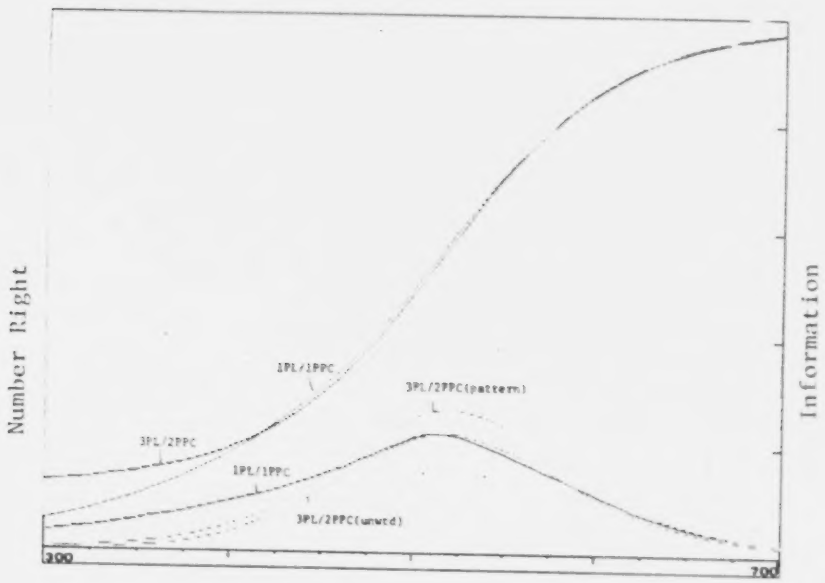




Scale Score	1pl/1ppc Raw Score	s.e.	3pl/2ppc(unwtd) Raw Score	s.e.
300	2.79	35.26	9.89	191.30
325	4.15	29.48	10.35	120.85
350	6.07	25.03	11.09	77.37
375	8.68	21.67	12.28	50.85
400	12.09	19.17	14.12	34.79
425	16.35	17.35	16.87	25.05
450	21.41	16.10	20.78	19.15
475	27.13	15.36	25.96	15.68
500	33.22	15.16	32.20	13.92
525	39.20	15.64	38.81	13.63
550	44.57	16.89	44.80	14.66
575	49.01	18.93	49.59	16.98
600	52.43	21.90	53.06	20.73
625	54.92	26.02	55.42	25.82
650	56.65	31.52	57.00	32.30
675	57.82	38.67	58.04	40.59
700	58.59	47.77	58.72	51.20

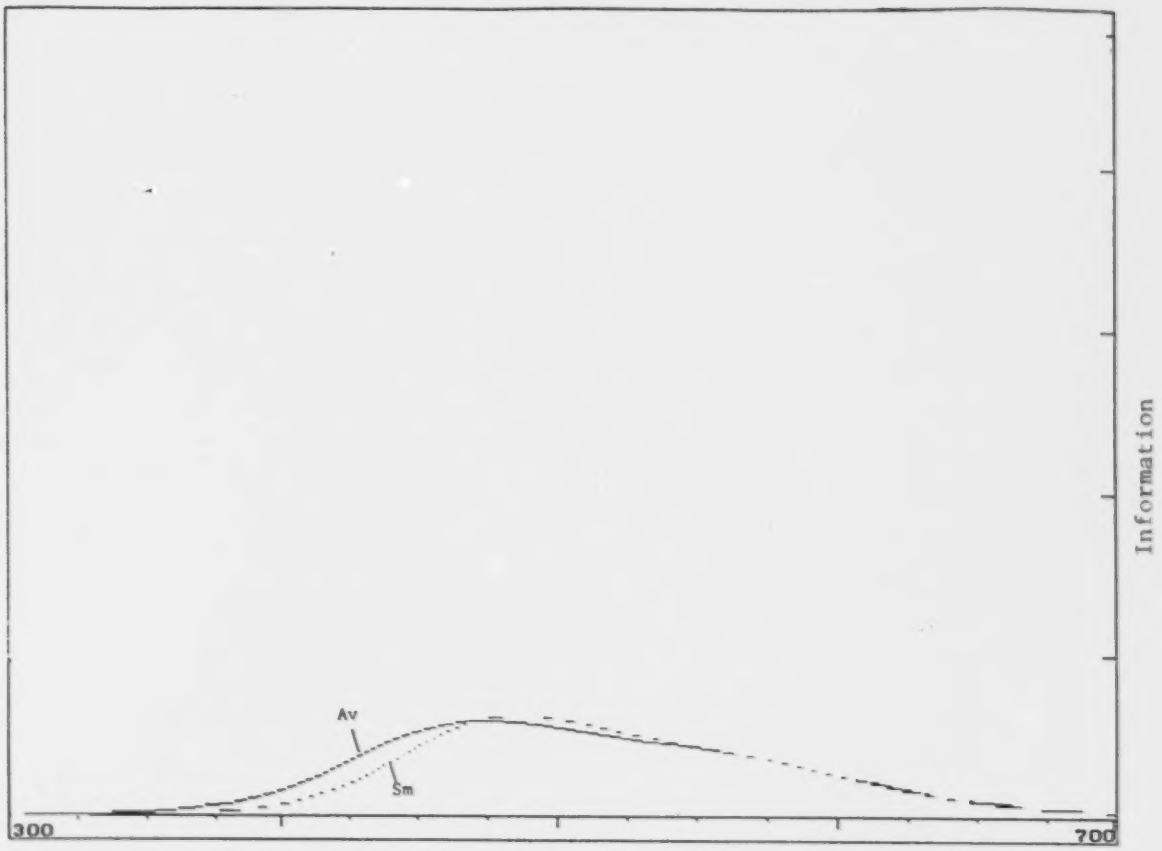
Figure 3b
Science: Summed Scores





Scale Score	1pl/1ppc Raw Score	s.e.	3pl/2ppc(unwtd) Raw Score	s.e.
300	4.51	34.74	10.13	163.40
325	6.14	30.37	10.68	106.08
350	8.27	26.78	11.55	69.80
375	10.97	23.80	12.92	47.18
400	14.38	21.28	15.06	33.17
425	18.62	19.12	18.27	24.44
450	23.85	17.22	22.94	19.04
475	30.32	15.48	29.34	15.84
500	38.24	14.14	37.41	14.20
525	47.05	14.08	46.34	13.96
550	55.09	15.33	54.73	14.90
575	61.68	17.11	61.68	16.86
600	66.81	19.85	66.93	20.20
625	70.45	24.05	70.58	25.04
650	72.89	29.54	73.03	31.14
675	74.51	36.12	74.66	38.63
700	75.60	43.89	75.73	47.87

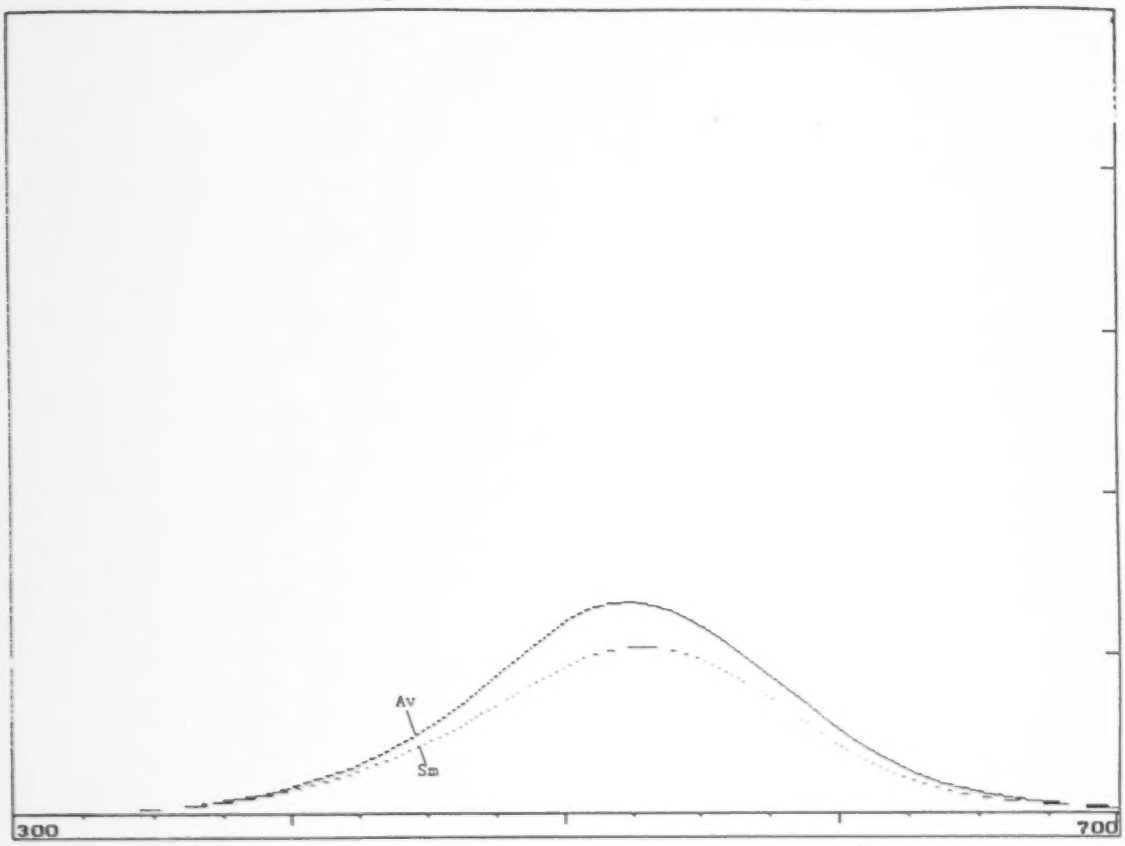
Figure 4
Reading: Averaged (Av) vs Summed (Sm) Ratings



Scale Score	Average Raw Score	Average Unwtd. s.e.	Sum Raw Score	Sum Unwtd. s.e.
300	7.31	211.50	7.90	335.58
325	7.66	136.94	8.13	216.44
350	8.22	88.35	8.50	137.94
375	9.12	56.70	9.09	86.54
400	10.58	36.40	10.08	52.99
425	12.92	24.86	11.77	32.36
450	16.18	20.10	14.55	22.13
475	19.89	19.03	18.32	18.78
500	23.52	19.81	22.45	18.84
525	26.76	21.16	26.40	20.42
550	29.61	22.48	30.01	22.24
575	32.02	24.75	33.21	24.74
600	33.89	28.99	35.74	29.16
625	35.21	36.06	37.51	36.78
650	36.09	46.49	38.63	48.08
675	36.65	59.99	39.33	62.24
700	37.02	76.34	39.78	78.62

Figure 5

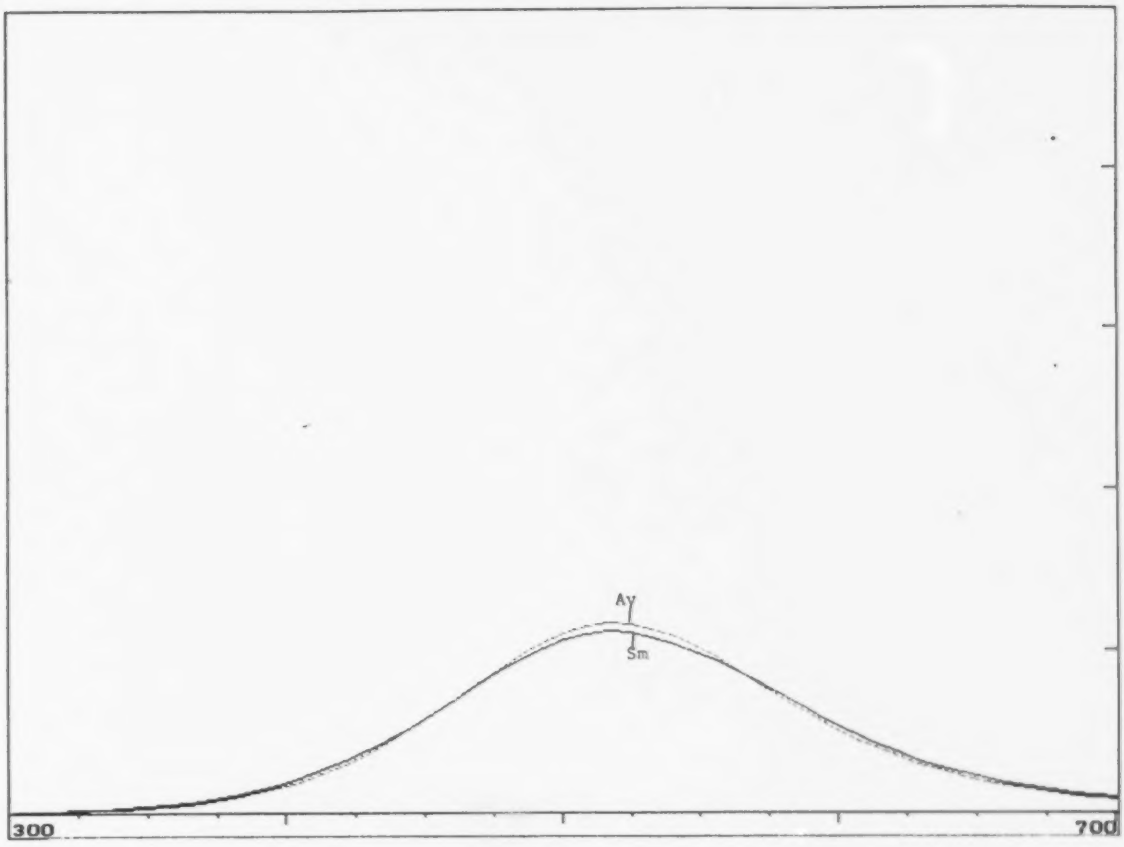
Math: Averaged (Av) vs Summed (Sm) Ratings



Scale Score	Average Raw Score	Average Unwtd. s.e.	Sum Raw Score	Sum Unwtd. s.e.
300	9.30	208.32	9.46	199.87
325	9.72	129.28	9.89	128.11
350	10.39	80.74	10.59	82.05
375	11.49	51.99	11.71	53.80
400	13.21	35.58	13.48	37.32
425	15.73	26.07	16.14	27.80
450	19.26	19.86	20.01	21.75
475	24.13	15.68	25.68	17.67
500	30.56	13.36	33.85	15.22
525	38.13	12.74	44.59	14.33
550	45.66	13.59	56.47	14.96
575	51.81	15.87	66.84	17.35
600	55.92	20.41	73.83	22.57
625	58.21	28.74	77.56	32.10
650	59.37	41.58	79.37	46.13
675	59.97	59.25	80.28	64.06
700	60.30	81.94	80.79	85.15

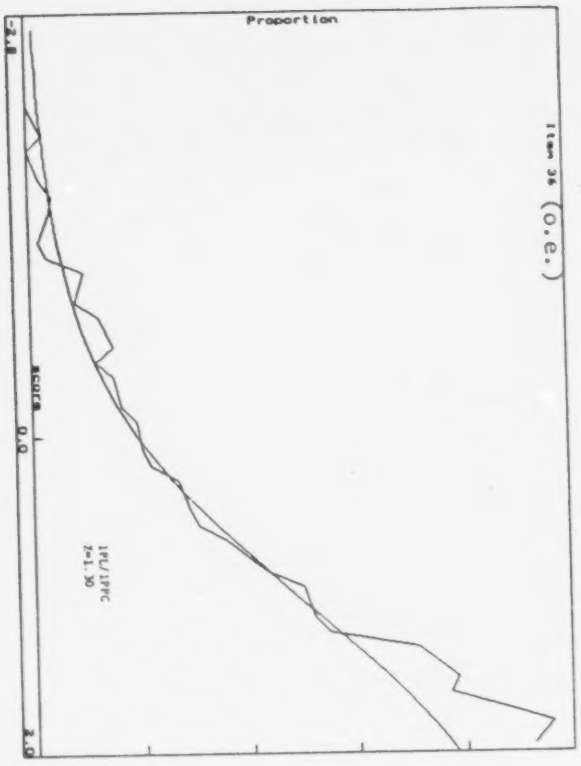
Figure 6

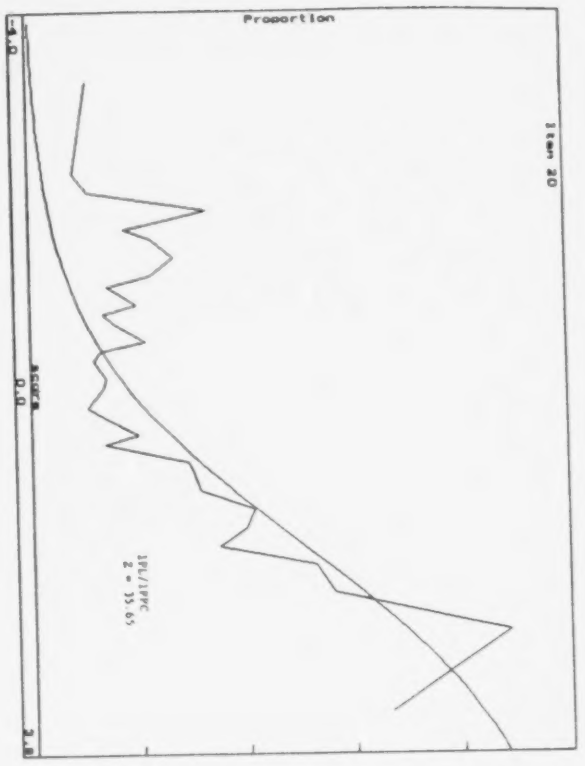
Science: Averaged (AV) vs Summed (Sm) Ratings

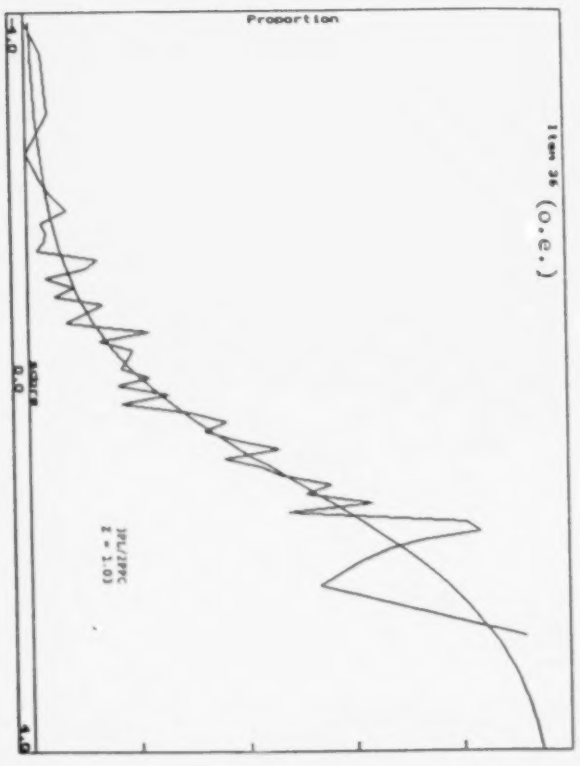


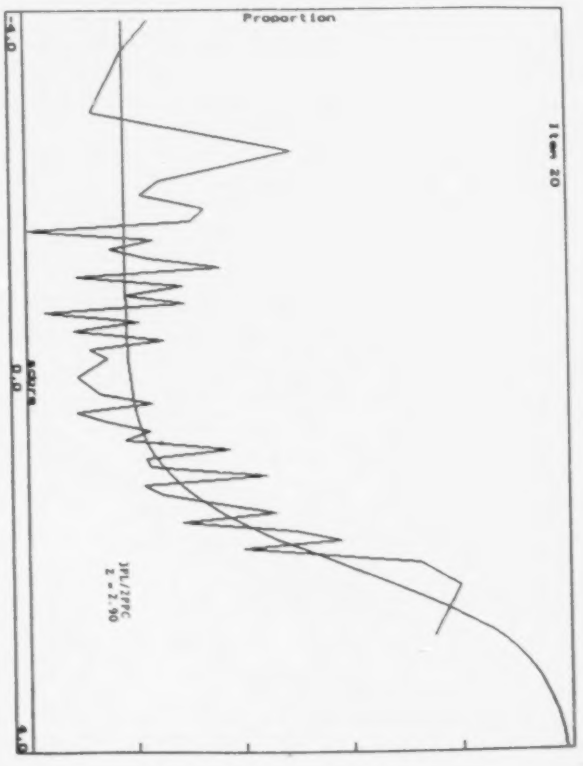
Scale Score	Average Raw Score	Average Unwtd. s.e.	Sum Raw Score	Sum Unwtd. s.e.
300	9.89	191.30	10.13	163.40
325	10.35	120.85	10.68	106.08
350	11.09	77.37	11.55	69.80
375	12.28	50.85	12.92	47.18
400	14.12	34.79	15.06	33.17
425	16.87	25.05	18.27	24.44
450	20.78	19.15	22.94	19.04
475	25.96	15.68	29.34	15.84
500	32.20	13.92	37.41	14.20
525	38.81	13.63	46.34	13.96
550	44.80	14.66	54.73	14.90
575	49.59	16.98	61.68	16.86
600	53.06	20.73	66.93	20.20
625	55.42	25.82	70.58	25.04
650	57.00	32.30	73.03	31.14
675	58.04	40.59	74.66	38.63
700	58.72	51.20	75.73	47.87

Appendix A1
Reading: Averaged Ratings

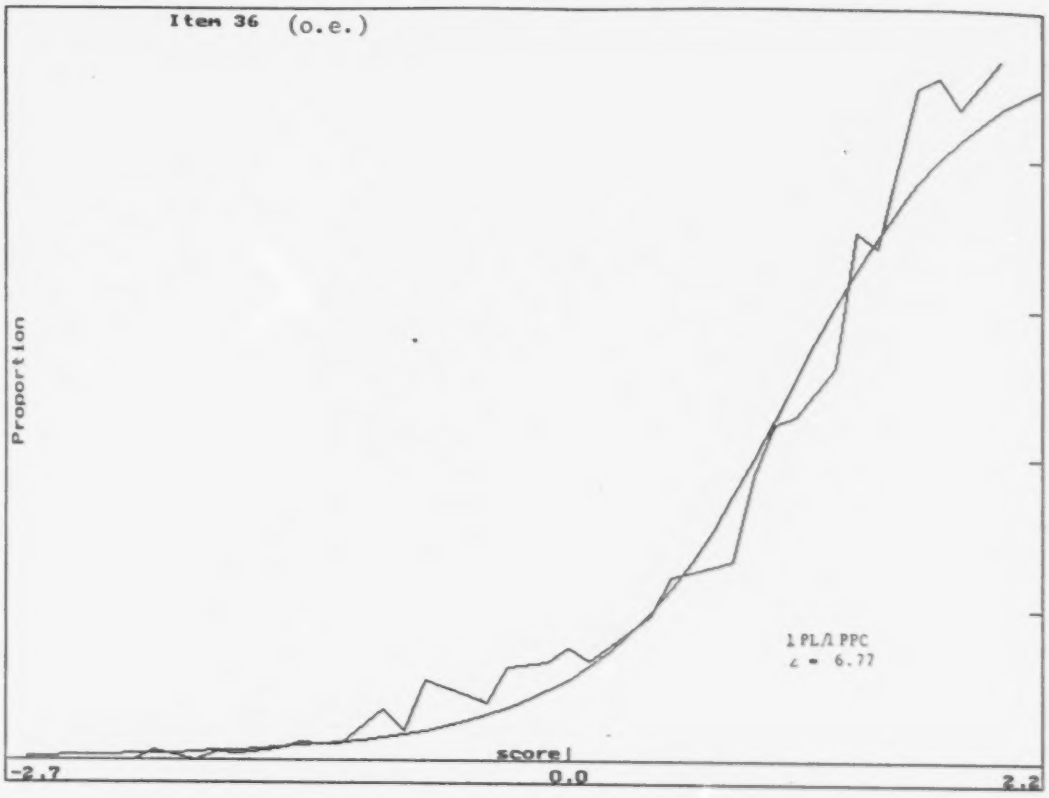


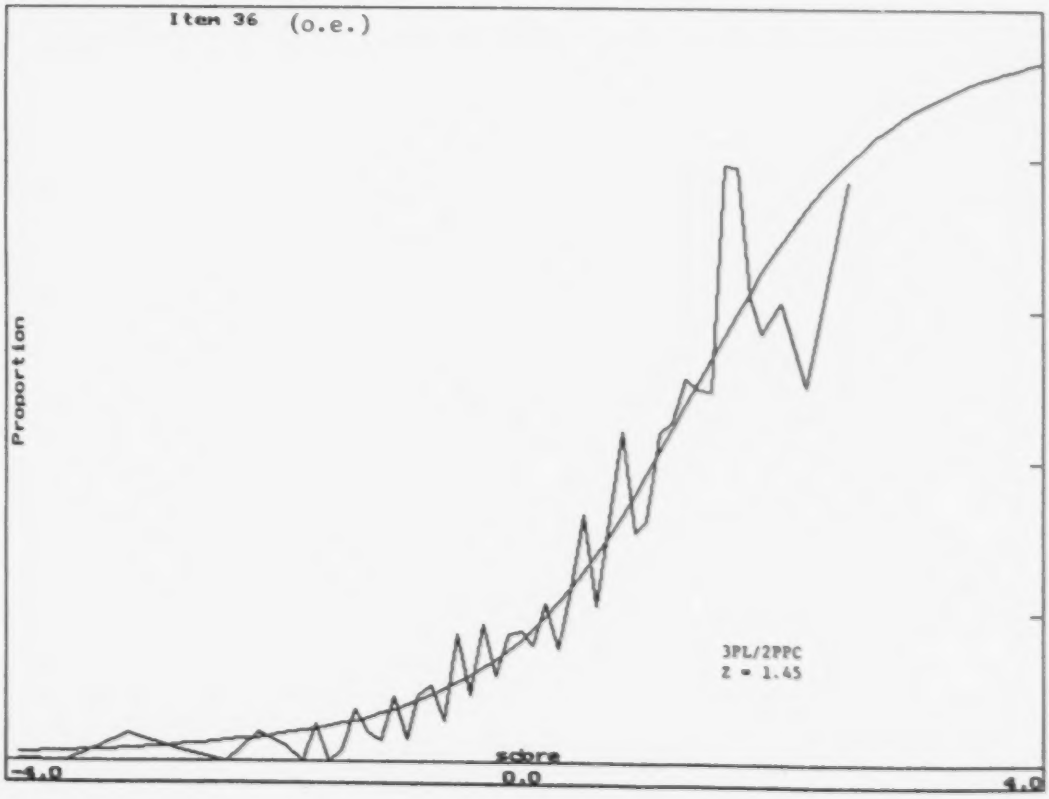




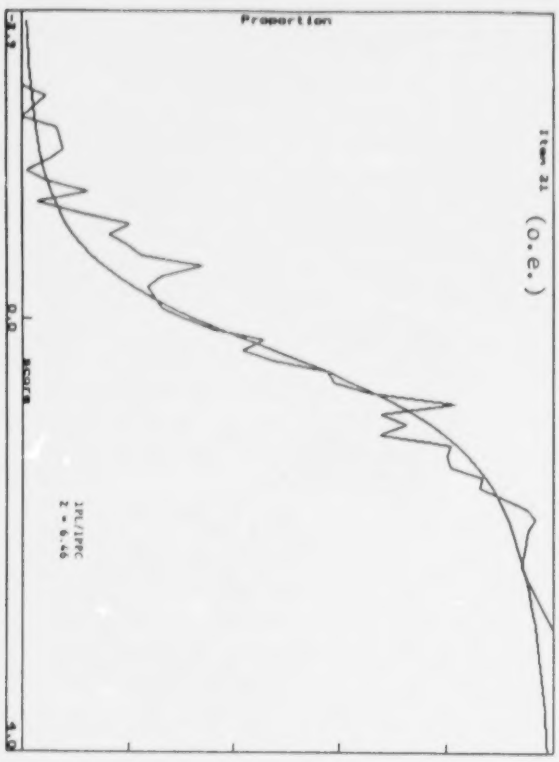


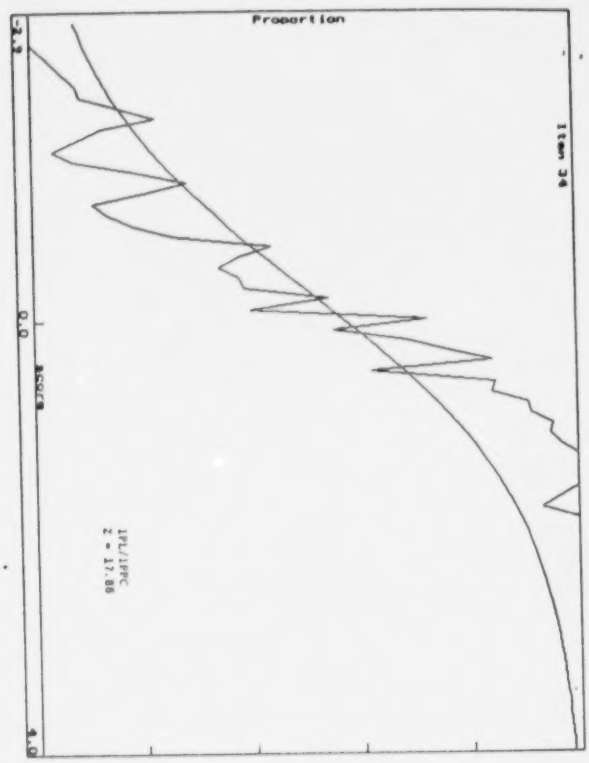
Appendix A2
Reading: Summed Ratings

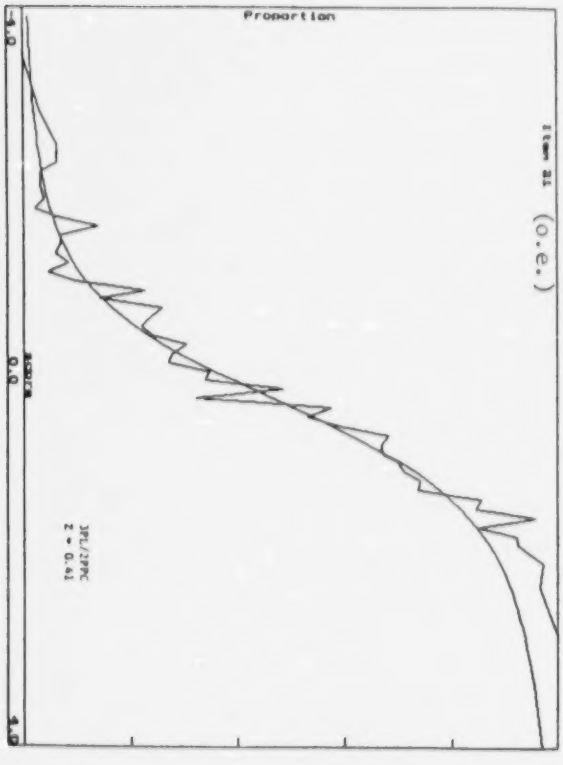


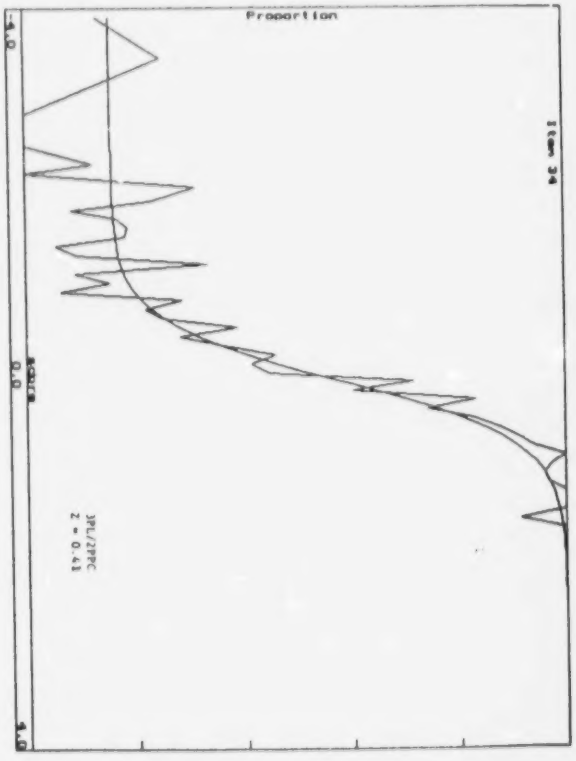


Appendix B1
Math: Averaged Ratings

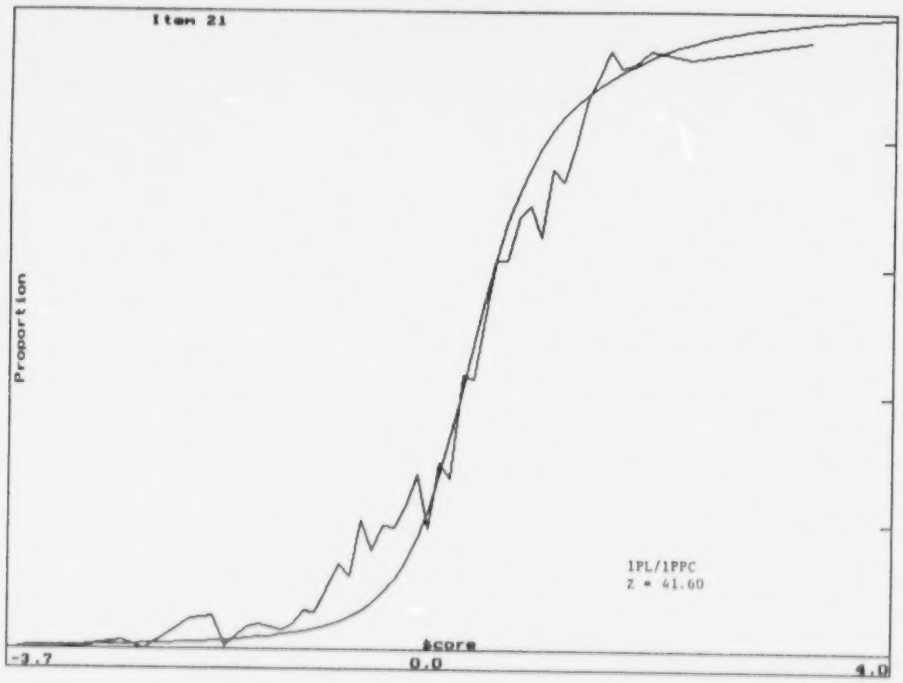


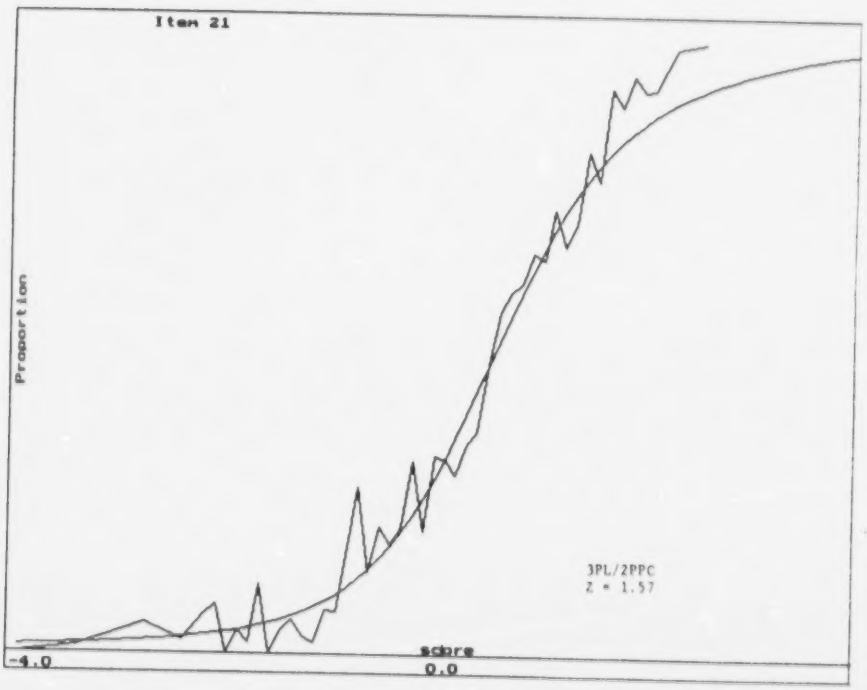




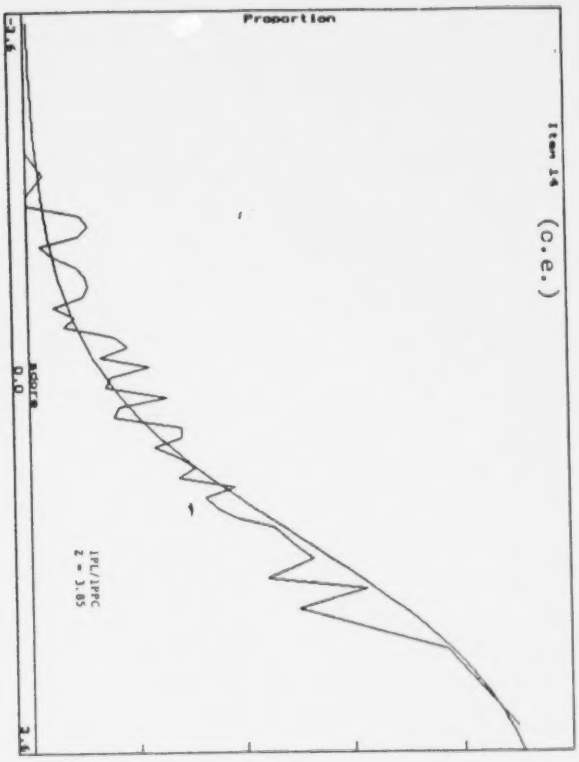


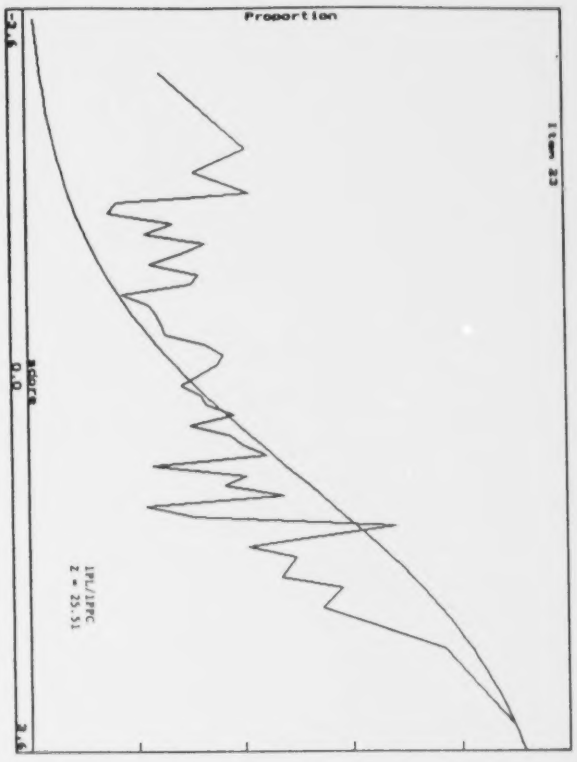
Appendix B2
Math: Summed Ratings

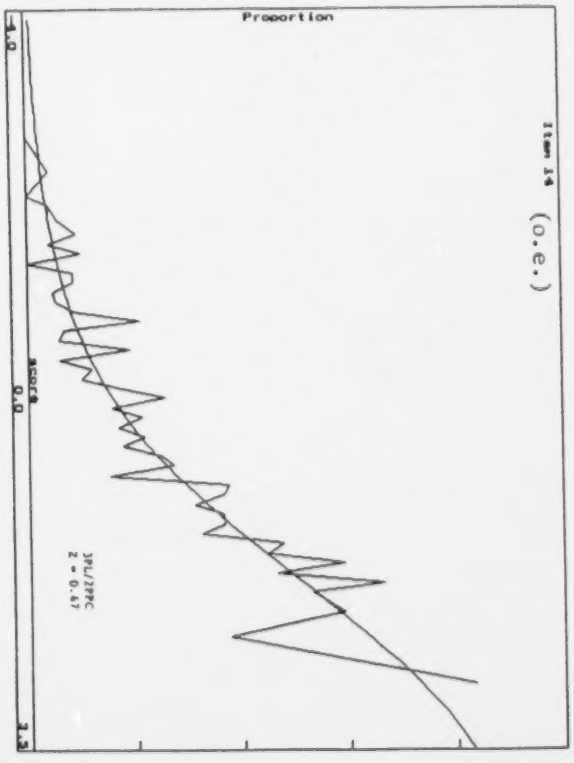


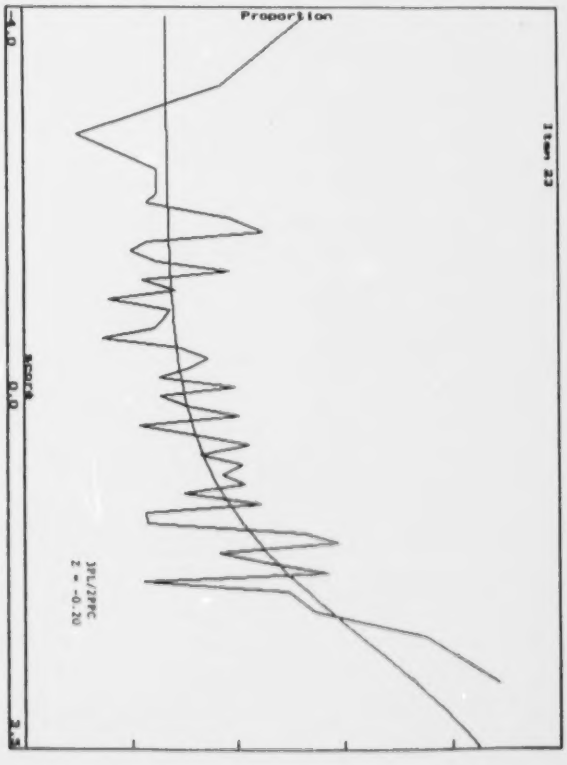


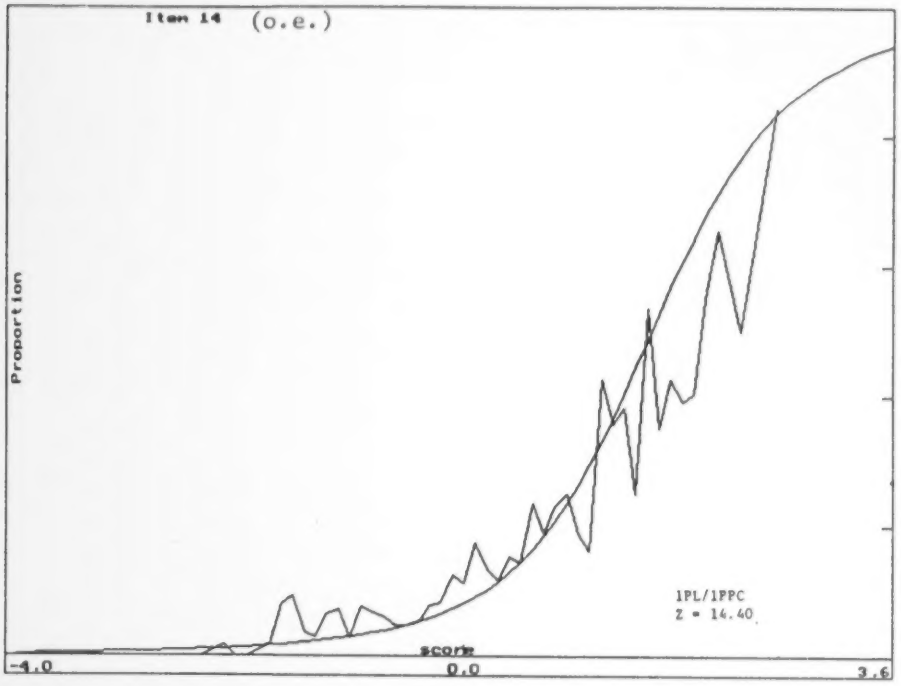
Appendix C1
Science: Averaged Ratings

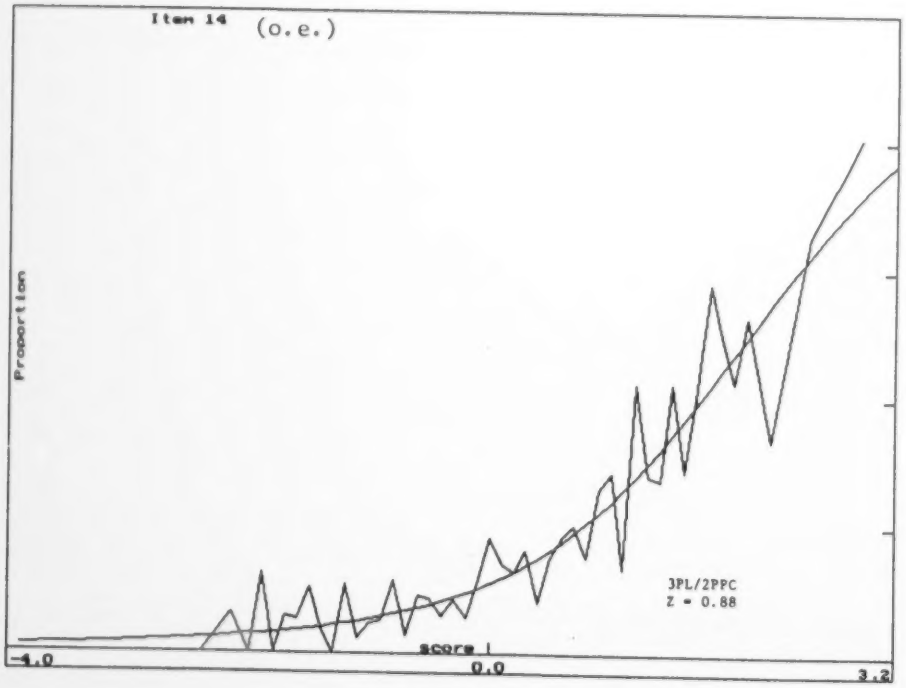









Item 14 (o.e.)
Proportion
1PL/1PPC
Z = 14.40
score
-4.0
0.0
3.6


Item 14 (o.e.)
Proportion
3PL/2PPC
Z = 0.88
score
-4.0
0.0
3.2

Item Parameters
Appendix D
Reading

	1PL/1PPC											3PL/2PPC										
	Average				Sum							Average				Sum						
Item	α	γ1	γ2	γ3	α	γ1	γ2	γ3	γ4	γ5	γ6	α	γ1	C (γ2)	γ3	α	γ1	C (γ2)	γ3	γ4	γ5	γ6
1	0.90	0.31			0.86	0.31						1.27	1.50	0.25		1.27	1.48	0.25				
2	0.90	0.13			0.86	0.12						0.98	0.80	0.20		0.98	0.80	0.20				
3	0.90	-0.84			0.86	-0.83						1.70	-0.34	0.24		1.74	-0.34	0.24				
4	0.90	-1.64			0.86	-1.63						2.10	-1.86	0.20		2.28	-1.82	0.25				
5	0.90	-1.17			0.86	-1.16						1.10	-0.82	0.20		1.11	-0.83	0.20				
6	0.90	-1.72			0.86	-1.71						1.15	-1.51	0.20		1.15	-1.51	0.20				
7	0.90	0.24			0.86	0.23						1.15	1.12	0.22		1.16	1.12	0.22				
8	0.90	0.38			0.86	0.37						0.87	1.12	0.20		0.88	1.12	0.20				
9	0.90	-1.46			0.86	-1.45						1.20	-1.21	0.20		1.20	-1.22	0.20				
10	0.90	-1.41			0.86	-1.40						1.00	-1.10	0.20		1.01	-1.10	0.20				
11	0.90	0.20			0.86	0.19						2.20	2.67	0.34		2.35	2.81	0.34				
12	0.90	-0.63			0.86	-0.63						1.41	-0.20	0.20		1.41	-0.21	0.20				
13	0.90	-0.79			0.86	-0.78						0.66	-0.34	0.20		0.66	-0.35	0.20				
14	0.90	0.39			0.86	0.38						0.77	1.41	0.25		0.74	1.35	0.25				
15	0.90	-0.54			0.86	-0.54						1.41	-0.05	0.20		1.42	-0.06	0.20				
16	0.90	-2.07			0.86	-2.06						2.76	-3.11	0.20		2.75	-3.11	0.20				
17	0.90	-0.77			0.86	-0.77						0.98	-0.35	0.20		0.99	-0.36	0.20				
18	0.90	-0.82			0.86	-0.82						0.80	-0.39	0.20		0.81	-0.40	0.20				
19	0.90	0.72			0.86	0.71						2.20	2.62	0.21		2.21	2.61	0.21				
20	0.90	1.41			0.86	1.39						2.18	4.35	0.19		2.30	4.46	0.19				
21	0.90	-1.34			0.86	-1.33						2.56	-1.10	0.30		2.59	-1.10	0.31				
22	0.90	-0.78			0.86	-0.77						1.76	-0.47	0.20		1.76	-0.48	0.20				
23	0.90	0.35			0.86	0.34						1.18	1.18	0.20		1.13	1.11	0.19				
24	0.90	-0.39			0.86	-0.39						1.19	0.12	0.20		1.20	0.12	0.20				
25	0.90	-1.09			0.86	-1.08						1.83	-0.91	0.20		1.83	-0.92	0.20				
26	0.90	-1.69			0.86	-1.68						3.00	-2.48	0.21		3.00	-2.49	0.20				
27	0.90	-0.09			0.86	-0.09						1.19	0.54	0.20		1.20	0.54	0.20				
28	0.90	-0.87			0.86	-0.87						1.57	-0.55	0.20		1.57	-0.56	0.20				
29	0.90	-0.25			0.86	-0.25						0.69	0.27	0.20		0.69	0.27	0.20				
30	0.90	-0.35			0.86	-0.35						2.06	0.14	0.17		2.05	0.12	0.17				
31	0.90	0.49			0.86	0.48						0.37	1.14	0.20		0.38	1.14	0.20				
32	0.90	-0.99			0.86	-0.98						1.71	-0.73	0.20		1.72	-0.74	0.20				
33	0.90	-0.46			0.86	-0.46						1.53	0.05	0.19		1.54	0.04	0.19				
34	0.90	-0.30			0.86	-0.30						1.66	0.54	0.24		1.68	0.53	0.24				
35	0.90	-0.78			0.86	-0.78						2.20	-0.49	0.19		2.21	-0.51	0.19				
36	0.90	0.68	0.84	1.23	0.86	1.41	0.10	0.89	1.01	0.92	0.96	0.74	0.72	(0.74)	1.02	0.39	1.55	(-0.04)	0.58	0.54	0.33	0.24

Item Parameters
Appendix D
Reading

	1PL/1PPC											3PL/2PPC										
	Average				Sum							Average				Sum						
Item	α	γ1	γ2	γ3	α	γ1	γ2	γ3	γ4	γ5	γ6	α	γ1	C (γ2)	γ3	α	γ1	C (γ2)	γ3	γ4	γ5	γ6
1	0.90	0.31			0.86	0.31						1.27	1.50	0.25		1.27	1.48	0.25				
2	0.90	0.13			0.86	0.12						0.98	0.80	0.20		0.98	0.80	0.20				
3	0.90	-0.84			0.86	-0.83						1.70	-0.34	0.24		1.74	-0.34	0.24				
4	0.90	-1.64			0.86	-1.63						2.10	-1.86	0.20		2.28	-1.82	0.25				
5	0.90	-1.17			0.86	-1.16						1.10	-0.82	0.20		1.11	-0.83	0.20				
6	0.90	-1.72			0.86	-1.71						1.15	-1.51	0.20		1.15	-1.51	0.20				
7	0.90	0.24			0.86	0.23						1.15	1.12	0.22		1.16	1.12	0.22				
8	0.90	0.38			0.86	0.37						0.87	1.12	0.20		0.88	1.12	0.20				
9	0.90	-1.46			0.86	-1.45						1.20	-1.21	0.20		1.20	-1.22	0.20				
10	0.90	-1.41			0.86	-1.40						1.00	-1.10	0.20		1.01	-1.10	0.20				
11	0.90	0.20			0.86	0.19						2.20	2.67	0.34		2.35	2.81	0.34				
12	0.90	-0.63			0.86	-0.63						1.41	-0.20	0.20		1.41	-0.21	0.20				
13	0.90	-0.79			0.86	-0.78						0.66	-0.34	0.20		0.66	-0.35	0.20				
14	0.90	0.39			0.86	0.38						0.77	1.41	0.25		0.74	1.35	0.25				
15	0.90	-0.54			0.86	-0.54						1.41	-0.05	0.20		1.42	-0.06	0.20				
16	0.90	-2.07			0.86	-2.06						2.76	-3.11	0.20		2.75	-3.11	0.20				
17	0.90	-0.77			0.86	-0.77						0.98	-0.35	0.20		0.99	-0.36	0.20				
18	0.90	-0.82			0.86	-0.82						0.80	-0.39	0.20		0.81	-0.40	0.20				
19	0.90	0.72			0.86	0.71						2.20	2.62	0.21		2.21	2.61	0.21				
20	0.90	1.41			0.86	1.39						2.18	4.35	0.19		2.30	4.46	0.19				
21	0.90	-1.34			0.86	-1.33						2.56	-1.10	0.30		2.59	-1.10	0.31				
22	0.90	-0.78			0.86	-0.77						1.76	-0.47	0.20		1.76	-0.48	0.20				
23	0.90	0.35			0.86	0.34						1.18	1.18	0.20		1.13	1.11	0.19				
24	0.90	-0.39			0.86	-0.39						1.19	0.12	0.20		1.20	0.12	0.20				
25	0.90	-1.09			0.86	-1.08						1.83	-0.91	0.20		1.83	-0.92	0.20				
26	0.90	-1.69			0.86	-1.68						3.00	-2.48	0.21		3.00	-2.49	0.20				
27	0.90	-0.09			0.86	-0.09						1.19	0.54	0.20		1.20	0.54	0.20				
28	0.90	-0.87			0.86	-0.87						1.57	-0.55	0.20		1.57	-0.56	0.20				
29	0.90	-0.25			0.86	-0.25						0.69	0.27	0.20		0.69	0.27	0.20				
30	0.90	-0.35			0.86	-0.35						2.06	0.14	0.17		2.05	0.12	0.17				
31	0.90	0.49			0.86	0.48						0.37	1.14	0.20		0.38	1.14	0.20				
32	0.90	-0.99			0.86	-0.98						1.71	-0.73	0.20		1.72	-0.74	0.20				
33	0.90	-0.46			0.86	-0.46						1.53	0.05	0.19		1.54	0.04	0.19				
34	0.90	-0.30			0.86	-0.30						1.66	0.54	0.24		1.68	0.53	0.24				
35	0.90	-0.78			0.86	-0.78						2.20	-0.49	0.19		2.21	-0.51	0.19				
36	0.90	0.68	0.84	1.23	0.86	1.41	0.10	0.89	1.01	0.92	0.96	0.74	0.72	(0.74)	1.02	0.39	1.55	(-0.04)	0.56	0.54	0.33	0.24

Item Parameters
Appendix E
Math

1PL/1PPC

Item	Average α	Average γ1	Average γ2	Average γ3	Average γ4	Sum α	Sum γ1	Sum γ2	Sum γ3	Sum γ4	Sum γ5	Sum γ6	Sum γ7	Sum γ8
1	0.97	-0.98				0.73	-0.95							
2	0.97	-1.78				0.73	-1.71							
3	0.97	-2.15				0.73	-2.07							
4	0.97	-2.45				0.73	-2.36							
5	0.97	-2.44				0.73	-2.35							
6	0.97	-1.38				0.73	-1.33							
7	0.97	-2.36				0.73	-2.28							
8	0.97	-0.71				0.73	-0.69							
9	0.97	-0.31				0.73	-0.31							
10	0.97	2.13				0.73	1.98							
11	0.97	-1.40				0.73	-1.35							
12	0.97	-0.59				0.73	-0.58							
13	0.97	-0.35				0.73	-0.35							
14	0.97	-0.64				0.73	-0.63							
15	0.97	-2.64				0.73	-2.56							
16	0.97	0.66				0.73	0.60							
17	0.97	-0.92				0.73	-0.89							
18	0.97	-0.96				0.73	-0.92							
19	0.97	0.15				0.73	0.12							
20	0.97	-0.24				0.73	-0.25							
21	0.97	0.57	0.63	-0.70	0.75	0.73	0.99	0.84	0.19	-0.06	-0.03	-0.68	1.34	-0.02
22	0.97	-0.60	-0.09	1.62	1.89	0.73	0.40	-0.83	0.86	-0.92	1.86	1.42	1.38	-0.00
23	0.97	1.62	0.82	-0.22		0.73	2.09	1.23	0.65	0.51	-0.23	-0.82		
24	0.97	0.60	0.83	0.35		0.73	2.65	-2.05	3.09	-1.96	3.05	-2.24		
25	0.97	2.45	1.28	0.41	-0.29	0.73	3.08	0.92	1.13	-0.53	0.96	1.01	0.68	-1.61
26	0.97	0.97	0.00			0.73	1.59	0.08	0.56	-0.43				
27	0.97	-1.33				0.73	-1.28							
28	0.97	0.14				0.73	0.11							
29	0.97	-0.94				0.73	-0.91							
30	0.97	0.46				0.73	0.41							
31	0.97	0.24				0.73	0.21							
32	0.97	-0.14				0.73	-0.15							
33	0.97	-0.16				0.73	-0.16							
34	0.97	-0.39				0.73	-0.38							
35	0.97	-1.07				0.73	-1.03							
36	0.97	-1.21				0.73	-1.17							
37	0.97	0.40				0.73	0.35							
38	0.97	-0.17				0.73	-0.18							
39	0.97	-0.26				0.73	-0.27							
40	0.97	-0.68				0.73	-0.68							
41	0.97	-0.48				0.73	-0.46							
42	0.97	-1.09				0.73	-1.05							
43	0.97	-0.00				0.73	-0.02							
44	0.97	0.91				0.73	0.84							
45	0.97	-1.62				0.73	-1.56							
46	0.97	0.26				0.73	0.22							

3PL/2PPC

Item	Average α	Average γ1	Average C (γ2)	Average γ3	Average γ4	Sum α	Sum γ1	Sum C (γ2)	Sum γ3	Sum γ4	Sum γ5	Sum γ6	Sum γ7	Sum γ8
1	1.07	-0.60	0.20			1.07	-0.59	0.20						
2	1.29	-1.66	0.20			1.28	-1.64	0.20						
3	1.87	-2.54	0.20			1.88	-2.53	0.20						
4	1.57	-2.71	0.20			1.56	-2.69	0.20						
5	0.89	-2.14	0.20			0.88	-2.13	0.20						
6	1.10	-1.08	0.20			1.10	-1.07	0.20						
7	1.27	-2.33	0.20			1.25	-2.31	0.20						
8	1.31	-0.07	0.26			1.32	-0.04	0.27						
9	2.14	1.29	0.34			2.22	1.37	0.35						
10	2.72	4.66	0.08			2.76	4.71	0.08						
11	1.28	-1.17	0.20			1.28	-1.16	0.20						
12	2.47	0.09	0.25			2.52	0.16	0.26						
13	1.98	0.63	0.27			1.98	0.66	0.27						
14	1.42	-0.15	0.22			1.45	-0.11	0.23						
15	1.81	-3.17	0.20			1.80	-3.15	0.20						
16	1.26	2.07	0.24			1.30	2.11	0.24						
17	1.70	-0.66	0.20			1.69	-0.64	0.20						
18	1.20	-0.60	0.20			1.20	-0.59	0.20						
19	1.33	1.22	0.25			1.43	1.33	0.26						
20	1.08	0.30	0.20			1.08	0.31	0.20						
21	0.64	0.78	(0.66)	-0.86	0.40	0.32	1.33	(0.96)	0.17	-0.19	-0.24	-0.97	0.97	-0.50
22	0.86	-0.52	(-0.09)	1.63	1.72	0.41	0.77	(-0.67)	0.84	-1.06	1.62	1.11	1.03	-0.40
23	0.91	1.63	(0.78)	-0.31		0.41	2.19	(1.14)	0.44	0.22	-0.57	-1.22	0.03	
24	1.08	0.56	(0.89)	0.45		0.53	2.79	(-2.06)	3.00	-2.11	2.86	-2.48	0.03	
25	0.59	2.46	(1.03)	-0.01	-0.86	0.28	3.16	(0.75)	0.82	-0.94	0.47	0.48	0.11	-2.23
26	1.20	0.93	(0.17)			0.70	1.62	(0.07)	0.53	-0.46				
27	0.87	-0.96	0.20			0.86	-0.94	0.20						
28	1.37	0.91	0.20			1.38	0.93	0.20						
29	0.91	-0.52	0.20			0.92	-0.51	0.20						
30	1.36	1.36	0.19			1.33	1.36	0.19						
31	0.38	0.83	0.20			0.38	0.83	0.20						
32	1.56	0.45	0.18			1.57	0.47	0.18						
33	1.79	0.63	0.22			1.83	0.69	0.22						
34	2.99	-0.03	0.16			3.00	0.03	0.16						
35	1.96	-0.61	0.29			2.00	-0.57	0.29						
36	2.53	-1.49	0.18			2.51	-1.44	0.19						
37	2.66	1.56	0.16			2.66	1.59	0.16						
38	1.66	1.11	0.31			1.69	1.17	0.31						
39	1.52	0.49	0.23			1.56	0.53	0.24						
40	1.91	-0.45	0.18			1.93	-0.40	0.18						
41	1.82	0.45	0.28			1.81	0.47	0.28						
42	1.77	-0.94	0.20			1.79	-0.92	0.20						
43	1.63	0.83	0.21			1.66	0.87	0.22						
44	2.50	3.57	0.22			2.67	3.78	0.22						
45	1.56	-1.60	0.20			1.55	-1.58	0.20						
46	1.82	1.34	0.22			1.66	1.40	0.23						

Item Parameters
Appendix F
Science

	1PL/1PPC											3PL/2PPC										
	Average				Sum							Average				Sum						
Item	α	γ1	γ2	γ3	α	γ1	γ2	γ3	γ4	γ5	γ6	α	γ1	C (γ2)	γ3	α	γ1	C (γ2)	γ3	γ4	γ5	γ6
1	0.89	-1.54			0.73	-1.49						0.89	-1.19	0.20		0.88	-1.19	0.20				
2	0.89	-1.38			0.73	-1.34						0.98	-1.04	0.20		0.99	-1.04	0.20				
3	0.89	-1.76			0.73	-1.70						1.15	-1.53	0.20		1.15	-1.53	0.20				
4	0.89	-1.87			0.73	-1.81						1.15	-1.68	0.20		1.13	-1.67	0.20				
5	0.89	-1.43			0.73	-1.38						1.34	-1.24	0.20		1.35	-1.24	0.20				
6	0.89	-0.39			0.73	-0.38						1.28	0.84	0.35		1.26	0.81	0.34				
7	0.89	-0.15			0.73	-0.15						0.86	0.43	0.20		0.84	0.43	0.20				
8	0.89	0.57			0.73	0.54						2.14	2.72	0.25		2.14	2.72	0.25				
9	0.89	0.68			0.73	0.65						0.86	1.81	0.22		0.87	1.84	0.23				
10	0.89	-0.01			0.73	-0.02						1.73	0.92	0.22		1.72	0.92	0.22				
11	0.89	-0.90			0.73	-0.87						1.18	-0.52	0.20		1.18	-0.51	0.20				
12	0.89	-0.87			0.73	-0.84						1.05	-0.48	0.20		1.05	-0.48	0.20				
13	0.89	-0.98			0.73	-0.95						1.23	-0.62	0.20		1.23	-0.62	0.20				
14	0.89	1.26	1.40		0.73	1.55	1.11	0.90	0.99			0.69	1.25	(1.16)		0.42	1.57	(0.91)	0.62	0.48		
15	0.89	-0.17	0.07		0.73	1.15	-1.28	0.93	-0.36			1.26	-0.32	(0.29)		0.69	1.22	(-1.24)	0.96	-0.38		
16	0.89	-0.02	0.12		0.73	1.12	-0.99	0.92	-0.32			1.27	-0.15	(0.38)		0.70	1.18	(-0.95)	0.93	-0.34		
17	0.89	-0.50			0.73	-0.48						1.38	0.02	0.20		1.37	0.02	0.20				
18	0.89	-0.62			0.73	-0.60						1.25	-0.18	0.20		1.24	-0.17	0.20				
19	0.89	-0.96			0.73	-0.93						0.88	-0.55	0.20		0.88	-0.54	0.20				
20	0.89	-1.21			0.73	-1.17						0.96	-0.83	0.20		0.94	-0.83	0.20				
21	0.89	-0.48			0.73	-0.47						1.61	0.13	0.22		1.57	0.12	0.22				
22	0.89	-0.72			0.73	-0.70						0.82	-0.26	0.20		0.81	-0.26	0.20				
23	0.89	0.79			0.73	0.76						1.21	2.86	0.27		1.11	2.70	0.26				
24	0.89	-0.23			0.73	-0.23						0.67	0.30	0.20		0.68	0.30	0.20				
25	0.89	-1.82			0.73	-1.76						2.02	-2.07	0.20		2.14	-2.03	0.24				
26	0.89	0.33			0.73	0.31						2.28	2.08	0.25		2.18	2.01	0.24				
27	0.89	-0.82			0.73	-0.79						0.86	-0.36	0.20		0.84	-0.35	0.20				
28	0.89	0.70			0.73	0.66						1.22	1.73	0.19		1.19	1.71	0.19				
29	0.89	-2.35			0.73	-2.28						1.52	-2.49	0.20		1.51	-2.49	0.20				
30	0.89	0.52	-0.20	1.54	0.73	1.26	-0.11	0.52	-0.90	1.96	0.64	1.07	0.48	(-0.09)	1.74	0.66	1.39	(-0.07)	0.49	-0.99	1.80	0.40
31	0.89	-1.00	0.13		0.73	0.35	-1.45	0.89	-0.24			1.35	-1.33	(0.34)		0.74	0.38	(-1.42)	0.93	-0.22		
32	0.89	1.67	-1.21		0.73	1.94	0.86	-0.93	-0.79			0.76	1.74	(-1.26)		0.41	2.11	(0.85)	-1.07	-1.07		
33	0.89	-0.78			0.73	-0.75						1.60	-0.42	0.20		1.57	-0.41	0.20				
34	0.89	0.52			0.73	0.49						0.86	1.33	0.20		0.84	1.33	0.20				
35	0.89	0.59			0.73	0.56						1.34	2.66	0.28		1.38	2.64	0.29				
36	0.89	-0.93			0.73	-0.90						1.14	-0.56	0.20		1.16	-0.54	0.20				
37	0.89	-1.09			0.73	-1.05						2.03	-0.73	0.25		2.00	-0.72	0.25				
38	0.89	0.33			0.73	0.31						0.77	1.02	0.20		0.77	1.02	0.20				
39	0.89	-0.20			0.73	-0.20						1.92	0.58	0.21		1.85	0.55	0.20				
40	0.89	0.34			0.73	0.31						1.26	1.20	0.20		1.25	1.20	0.20				
41	0.89	-0.28			0.73	-0.27						2.19	0.80	0.26		2.16	0.82	0.26				
42	0.89	-0.22			0.73	-0.22						0.51	0.29	0.20		0.51	0.30	0.20				
43	0.89	-0.76			0.73	-0.72						1.54	-0.34	0.20		1.52	-0.34	0.20				
44	0.89	-0.62			0.73	-0.60						1.12	-0.02	0.20		1.12	-0.01	0.20				
45	0.89	-0.28			0.73	-0.27						2.28	0.64	0.23		2.24	0.62	0.23				
46	0.89	0.39	1.17	1.56	0.73	1.60	-0.90	1.33	1.16	1.21	0.87	1.43	0.35	(1.60)	2.34	0.82	1.60	(-0.84)	1.42	1.28	1.36	1.04
47	0.89	-0.43			0.73	-0.42						0.90	0.09	0.20		0.90	0.10	0.20				
48	0.89	0.07			0.73	0.06						1.52	1.04	0.23		1.50	1.03	0.23				
49	0.89	-1.06			0.73	-1.02						1.88	-0.55	0.28		1.88	-0.55	0.28				
50	0.89	-1.22	-0.21		0.73	0.67	-2.22	0.37	0.20			0.93	-1.25	(-0.17)		0.51	0.96	(-2.07)	0.38	0.06		